LSA(Latent Semantic Analysis)
の概要と実行

Ｄ１　猪原敬介

心理データ解析演習_090624

アウトライン

- LSAのざっくりとした説明
- 小規模コーパスでの具体例
- 文献が語る性能
- 実際に大規模コーパスから作成した意味空間のデモ

LSAとは何か

- LSA(潜在意味解析; Latent Semantic Analysis; Landauer & Dumais, 1997)
  - 初めに情報検索の技術として提案される(Dumais, Furnas, Landauer, & Deerwester, 1988; Deerwester, Dumais, Furnas, Landauer, & Harshman, 1990)。
    - 情報検索における索引づけの技術なので，Latent Semantic Indexing (LSI)とも呼ばれる。
    - LSAは「ベクトル空間モデル」と呼ばれるモデル群の一部。
    - 同義語 (e.g., “car”と“automobile”) 間で別々の検索結果が出てくるという問題をクリアする。
  - 単語や文書の意味を学習し，ベクトルとして表現する。
  - 人間の語彙獲得・表象理論としても研究。

LSAの何が良いか

- 単語-単語，単語-文書，文書-文書の**類似性**を（人間の直感に近い値で）計算することができる。
  - 単語・文書の意味を学習するモデルであると言える。

- 大規模言語コーパス（後述）から，**人手の介入を一切なしに，**自動的な学習を行う。
  - 少なくとも計算機上で実行可能であることが，具体的なパフォーマンスと共に保障されている。
  - 少なくとも1997年当時，他にこれができる理論はなかった（Landauer & Dumais, 1997）。

LSAの何が良いか

- LSAベースの認知モデルを構築することで，想定したアルゴリズムがきちんと行動データを説明するのかどうかを，人間が持つ知識に近い大きさの知識モデルで検討できる。
  - これまでは人手の小さい知識モデルで，概念間のリンク強度は研究者の直感に基づいたものであった。
  - どのような概念間にも小さい類似度が存在するため，それらが集積して大きな影響力を持つことを小さいモデルでは無視していた。

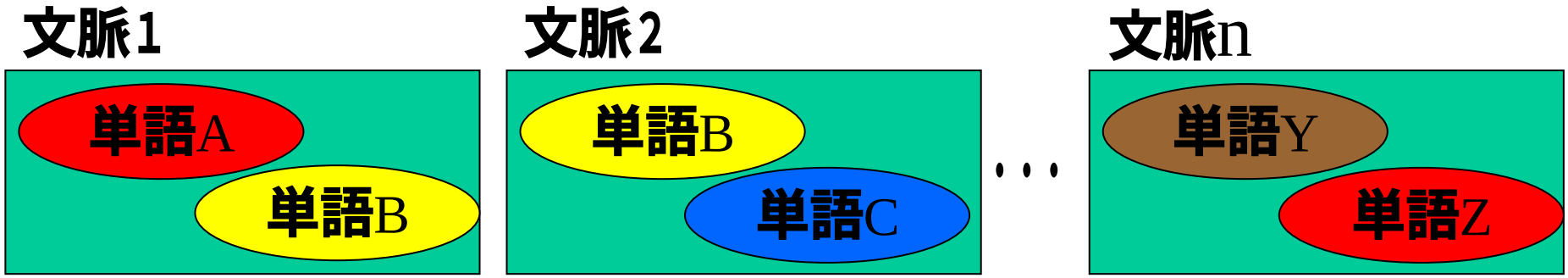
LSAの基本アイディア

- **基本的アイディア**
  - 類似した意味を持つものは，類似した用いられ方をするはず。
  - 類似した意味を持つものは，類似した文脈に共起するはず。
  - 人間も，もの‐文脈という共起関係から意味を学習するのでは。

- このアイディアを確かめるのに、コーパスを用いる。
- **大規模言語コーパス**
  - 電子化された教科書，新聞，百科事典など。
  - 実世界の学習環境に見立てられる大きさと構造。
  - 単語‐文，段落，文書の共起関係から、単語の意味を学習。
  - コーパスに基づいた計算機シミュレーションが行われ，少なくともある程度，上の主張の妥当性が支持されている。

LSAの直感的説明

- 文脈　＝　コーパスの文，段落，記事，文書
- 共起関係の例
  - N次的共起関係
- こうした共起パターンには，ノイズが含まれる。また，共起パターンに類似があっても非常に遠い関係しかないかもしれない。
- そこで，情報を削らないように何十万次元を数百次元に縮約して表現させると，ノイズが減り，遠くても安定した共起パターンが強調される。
- この強調されたパターンを単語の意味のベクトルとして保持する。



LSAの具体的手続き

小規模コーパスでの具体例

手続きのアウトライン

- コーパスを用意する
- 単語文脈行列
- 特異値分解
- 近似行列の再構築
- 類似度の計算

コーパスを用意する

- 通常は，新聞コーパスなどを用いる。
  - 1記事＝1文脈として，数十万文脈となる。

- 今回は説明のために，Landauer, Foltz, & Laham(1998)の小さいコーパスを用いる。
  - ある論文のタイトル9つをコーパスとする。
    - 5つは人間-機械コミュニケーション
    - 4つは数学のグラフ理論

コーパス例

Example of text data: Titles of Some Technical Memos

c1: *Human* machine *interface* for ABC *computer* applications
c2: A *survey* of *user* opinion of *computer system response time*
c3: The *EPS user interface* management *system*
c4: *System* and *human system* engineering testing of *EPS*
c5: Relation of *user* perceived *response time* to error measurement

m1: The generation of random, binary, ordered *trees*
m2: The intersection *graph* of paths in *trees*
m3: *Graph minors* IV: Widths of *trees* and well-quasi-ordering
m4: *Graph minors*: A *survey*

コーパス → 単語文脈行列

- 単語と文脈の共起パターンを表現する行列。

- 今回は，1タイトル＝1文脈とする。
- 文脈を列に，異なり単語（　種類）を行に配する。
  - ⇔ 延べ単語（　総数）
- 各セルはその文脈でその単語が登場した回数。

- ただし，今回の例では，単語は「2つ以上のタイトルで用いられたもの」に限定している。
  - 抽出された単語は前スライドでイタリックになっている。
  - 12単語

単語文脈行列

$\{X\} =$

	c1	c2	c3	c4	c5	m1	m2	m3	m4
human	1	0	0	1	0	0	0	0	0
interface	1	0	1	0	0	0	0	0	0
computer	1	1	0	0	0	0	0	0	0
user	0	1	1	0	1	0	0	0	0
system	0	1	1	2	0	0	0	0	0
response	0	1	0	0	1	0	0	0	0
time	0	1	0	0	1	0	0	0	0
EPS	0	0	1	1	0	0	0	0	0
survey	0	1	0	0	0	0	0	0	1
trees	0	0	0	0	0	1	1	1	0
graph	0	0	0	0	0	0	1	1	1
minors	0	0	0	0	0	0	0	1	1

r (human.user) = -.38

r (human.minors) = -.29

特異値分解を実行

- **単語文脈行列**Xを，W，S, P**の3つの行列**に分解する。
- 線形分解であり，丸めの誤差を除いて，分解された行列から元の行列を完全に復元できる。
  - W： 左特異ベクトル。単語を表現。
  - S： 特異値（ 重要度）が入った対角行列
  - P： 右特異ベクトル。文脈を表現。

$$\{X\} = \{W\}\{S\}\{P\}'$$

W：12×9

$$\{W\} =$$

0.22	-0.11	0.29	-0.41	-0.11	-0.34	0.52	-0.06	-0.41
0.20	-0.07	0.14	-0.55	0.28	0.50	-0.07	-0.01	-0.11
0.24	0.04	-0.16	-0.59	-0.11	-0.25	-0.30	0.06	0.49
0.40	0.06	-0.34	0.10	0.33	0.38	0.00	0.00	0.01
0.64	-0.17	0.36	0.33	-0.16	-0.21	-0.17	0.03	0.27
0.27	0.11	-0.43	0.07	0.08	-0.17	0.28	-0.02	-0.05
0.27	0.11	-0.43	0.07	0.08	-0.17	0.28	-0.02	-0.05
0.30	-0.14	0.33	0.19	0.11	0.27	0.03	-0.02	-0.17
0.21	0.27	-0.18	-0.03	-0.54	0.08	-0.47	-0.04	-0.58
0.01	0.49	0.23	0.03	0.59	-0.39	-0.29	0.25	-0.23
0.04	0.62	0.22	0.00	-0.07	0.11	0.16	-0.68	0.23
0.03	0.45	0.14	-0.01	-0.30	0.28	0.34	0.68	0.18

S：特異値(9)を対角成分に持つ行列

$$\{S\} = \begin{matrix} 3.34 & & & & & & & & \\ & 2.54 & & & & & & & \\ & & 2.35 & & & & & & \\ & & & 1.64 & & & & & \\ & & & & 1.50 & & & & \\ & & & & & 1.31 & & & \\ & & & & & & 0.85 & & \\ & & & & & & & 0.56 & \\ & & & & & & & & 0.36 \end{matrix}$$

P：9×9行列

$$\{P\} =$$

0.20	0.61	0.46	0.54	0.28	0.00	0.01	0.02	0.08
-0.06	0.17	-0.13	-0.23	0.11	0.19	0.44	0.62	0.53
0.11	-0.50	0.21	0.57	-0.51	0.10	0.19	0.25	0.08
-0.95	-0.03	0.04	0.27	0.15	0.02	0.02	0.01	-0.03
0.05	-0.21	0.38	-0.21	0.33	0.39	0.35	0.15	-0.60
-0.08	-0.26	0.72	-0.37	0.03	-0.30	-0.21	0.00	0.36
0.18	-0.43	-0.24	0.26	0.67	-0.34	-0.15	0.25	0.04
-0.01	0.05	0.01	-0.02	-0.06	0.45	-0.76	0.45	-0.07
-0.06	0.24	0.02	-0.08	-0.26	-0.62	0.02	0.52	-0.45

近似行列の作成

- 今回は，2次元に縮約する。
- Sの上位2つだけを用いて，行列を再構築する。
  - 黄色マーカーのついている部分だけで計算する。
  - Pは転置なので，(12,2)(2,2)(2,9)の積。

近似行列

$\{\hat{X}\} =$

	c1	c2	c3	c4	c5	m1	m2	m3	m4
human	0.16	0.40	0.38	0.47	0.18	-0.05	-0.12	-0.16	-0.09
interface	0.14	0.37	0.33	0.40	0.16	-0.03	-0.07	-0.10	-0.04
computer	0.15	0.51	0.36	0.41	0.24	0.02	0.06	0.09	0.12
user	0.26	0.84	0.61	0.70	0.39	0.03	0.08	0.12	0.19
system	0.45	1.23	1.05	1.27	0.56	-0.07	-0.15	-0.21	-0.05
response	0.16	0.58	0.38	0.42	0.28	0.06	0.13	0.19	0.22
time	0.16	0.58	0.38	0.42	0.28	0.06	0.13	0.19	0.22
EPS	0.22	0.55	0.51	0.63	0.24	-0.07	-0.14	-0.20	-0.11
survey	0.10	0.53	0.23	0.21	0.27	0.14	0.31	0.44	0.42
trees	-0.06	0.23	-0.14	-0.27	0.14	0.24	0.55	0.77	0.66
graph	-0.06	0.34	-0.15	-0.30	0.20	0.31	0.69	0.98	0.85
minors	-0.04	0.25	-0.10	-0.21	0.15	0.22	0.50	0.71	0.62

r (human.user) = .94

r (human.minors) = -.83

単語文脈行列 (再掲)

$\{X\} =$

	c1	c2	c3	c4	c5	m1	m2	m3	m4
human	1	0	0	1	0	0	0	0	0
interface	1	0	1	0	0	0	0	0	0
computer	1	1	0	0	0	0	0	0	0
user	0	1	1	0	1	0	0	0	0
system	0	1	1	2	0	0	0	0	0
response	0	1	0	0	1	0	0	0	0
time	0	1	0	0	1	0	0	0	0
EPS	0	0	1	1	0	0	0	0	0
survey	0	1	0	0	0	0	0	0	1
trees	0	0	0	0	0	1	1	1	0
graph	0	0	0	0	0	0	1	1	1
minors	0	0	0	0	0	0	0	1	1

r (human.user) = -.38

r (human.minors) = -.29

近似行列

- 元の単語文脈行列Xは、0の多いスパース行列であった。
- しかし、近似行列X^には、元々0であった部分にも値が入っている。

- 近似行列における各セルの値は、その単語がその文脈を表現するのにどれだけ貢献しているかを示している。

SVDの効果

- m4のSurveyとtreesに注目
  - m4：”Graph minors: A survey”
- m4にはtree（グラフ理論の用語）は出てこないにも関わらず，X^では.66となっている。
  - これは，m4がgraphとminorsを含んでいて，この2つとtreeは共起しやすいことが原因と考えられる。
  - 他の共起関係からm4を表現するのに重要なため。
- 反対に，surveyは.44と下がっている。
  - 他の共起関係からm4を特徴づけるのに重要ではなかったため。

Xにおける単語間類似性

- human，user，minorsに注目してみる。
  - XとX^を比較してみる。
  - Minorはグラフ理論の専門用語らしい。
- Xでは，humanと他の2つは共起していない。

- 相関係数を見てみる。
  - Human-user: -.38
  - Human-minor: -.29
  - 直感的にはhuman-userは正だが，そうなっていない。

X^における単語間類似性

- LSAによる近似行列 (X^) における相関係数
  - Human-user: .94
  - Human-minor: -.83
  - 直感に合うように変化している。
  - 単純な共起 (X) ではなく，背後の関係を反映。

文献が語るLSAの性能

LSAの性能とは

- やはり「人間にどれだけ似た振る舞いをするのか」によって定めるべき。

- シミュレーションと比較できるような心理学実験を考えて，人間のデータとシミュレーション結果を比較する研究が行われている。

シミュレーション例

- Word categorization (Laham, 2000)
- Discourse　comprehension (Kintsch, 1998)
- Judgments of essay quality (Landauer, Laham, Rehder, & Schreiner, 1997)
- **など**

- **以下を紹介**
- Judgments of semantic similarity (Landauer & Dumais, 1997)

Landauer & Dumais(1997)

- **同義語テストについてLSAと人間の比較**
  - ETS(Educational Testing Service)が行っているTOEFL(Test of English as a Foreign Language)
  - **ターゲット単語に対して4つの単語が選択肢としてあり，ターゲット単語に最も近い意味の単語を4つの中から選べというもの。**
- **人間**
  - **非英語母語者で，アメリカの大学を受験して合格した人の成績**

Landauer & Dumais(1997)

- **ＬＳＡ**
  - **ターゲット語-選択肢の類似度が一番高い選択肢を回答したことにする。**

- **学習したコーパス**
  - **標準的なアメリカの大学一年生がそれまでに読んできたテキストとほぼ同じサイズと内容のコーパスを用いた。**
  - K-12 (**アメリカの12年間教育。小中高と考えればよい**) **の学校にある本**
  - TASA(Touchstone Applied Science Associate)

Landauer & Dumais(1997)

- **LSAと合格者は同程度の成績であった。**
  - **８０問中，ＬＳＡは51.5問（６４．４％）に正解。非英語母語者で合格者は51.6問（６４．５％）に正解。**
- **コーパスサイズの増加とLSAの成績向上とを曲線にすると，読書量の増加とアメリカの子供の語彙獲得量との曲線と非常に良く似ている。**

- **ここから，LSAと人間の学習は，同じような振る舞いをしていると言うことができる。**