アクセス制御のための Boid によるテクスト依存関係の学習

木下研究室　201503789　紅林 宏祐

研究背景

- 近年，SNSにおける情報発信やビッグデータの解析などによりさまざまな恩恵が得られる反面，プライバシーの侵害が問題になっている.
- しかし論理学的命題において表現されるアクセス制御の表現には限界がある.
- したがって従来のアクセス制御の枠組みではあつかうことが困難であった推論攻撃による情報漏えいに対処する必要が出てきた.

目的

- そこで新しい概念のセキュリティモデルを目指し，確率論に着目した推論規則を生成する機械学習モデルを研究する.
- 推論攻撃は大量のテクストの確率的な振る舞いによって成立する現象である，と見做せる.
- 確率的な現象は情報理論的な評価尺度が必要になる.
- 本稿ではテクストの類似の振る舞いを確率変数とし，確率変数の確率分布を機械学習するモデルを提案する.

提案手法の位置付け

テクストの族に対して，word2vecを使用することによりPMI情報量

（前順序関係(pre-order),反射率および推移律）を導入し，

cohesion, separation を評価するKL情報量（完全律）の Boidを導入する。

この様に，情報理論的に設計したBoid（半順序関係）は機械学習のための前処理に位置づけられるモデルである。

提案手法

- 推論規則を学習するモデルの作成
- モデルの実装方法

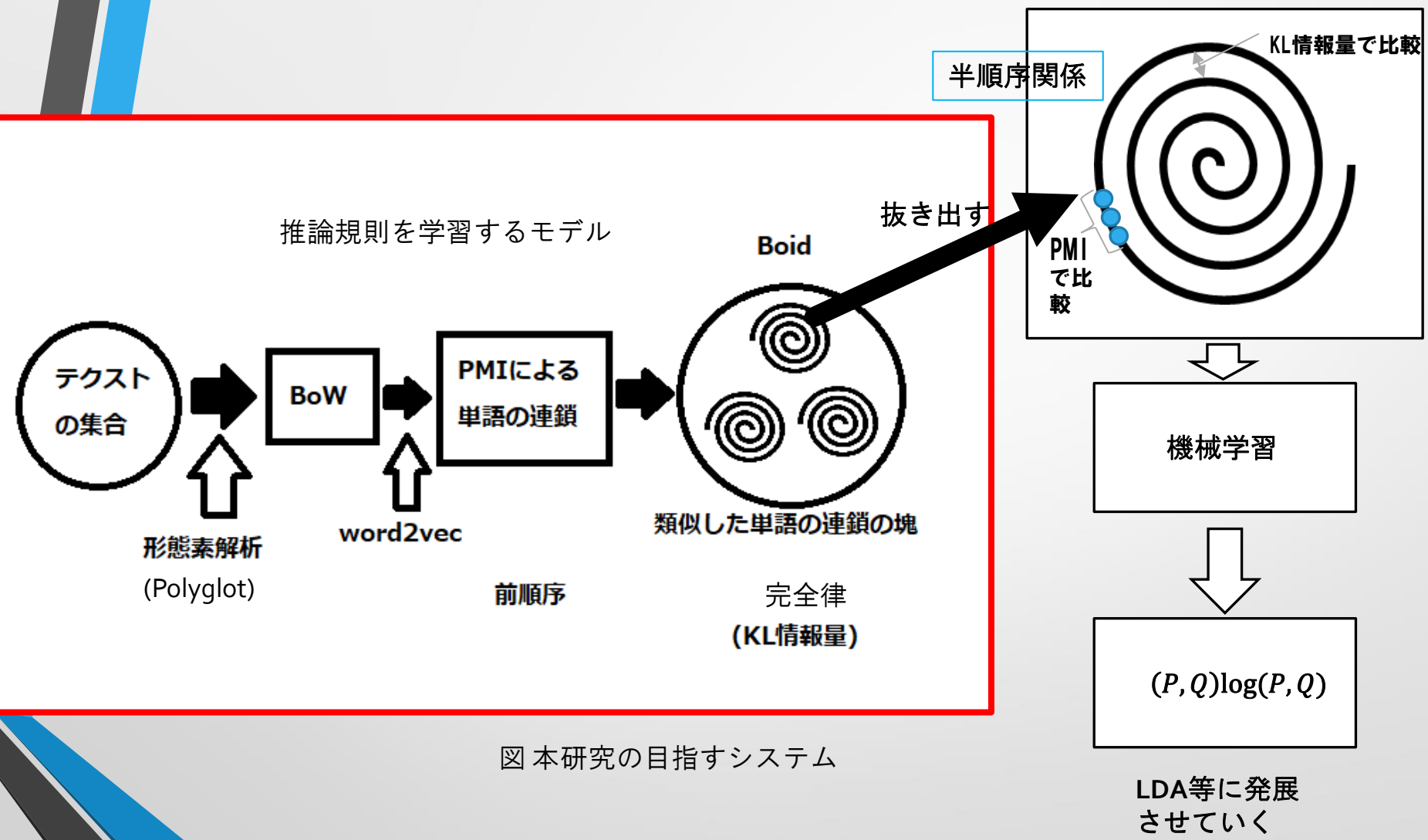
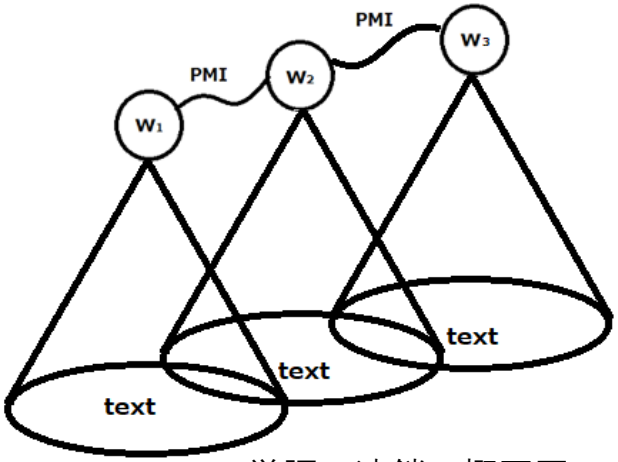


図 本研究の目指すシステム

テクストに含まれる互いに類似する単語の連鎖を作る

- 単語の連鎖を作るにあたり，word2vecを使用する。
- word2vecにはPMIという評価尺度がある。
- このPMIが前順序関係であり反射律，推移律だけを満たす。したがって，PMIで評価した単語の連鎖どうしは比較不可能である。（PMIは相関関係を示すものであるため）



単語の連鎖の概要図

KL情報量を導入する

- KL情報量は以下の式から求められる

$$D_{kl}\big(p(\theta|y)\big)||p(\theta)\big) = \sum_{\theta} p(\theta|y) log \frac{p(\theta|y)}{p(\theta)}$$

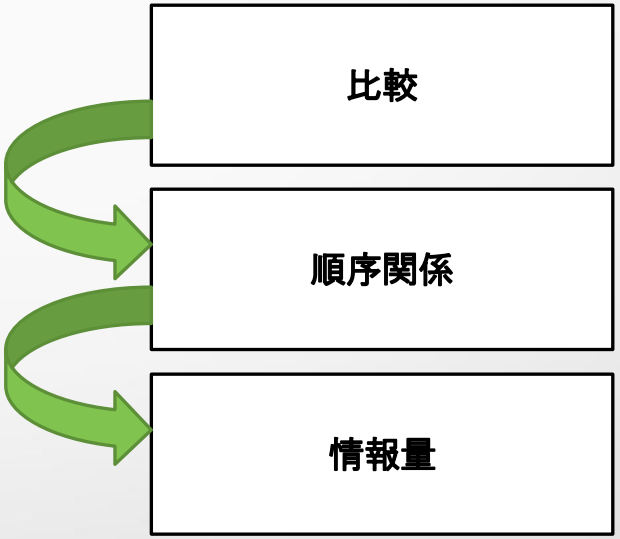
- KL情報量は完全律を満たす。すなわち比較可能である。
- しかし，対称律，反対称律，推移律を満たさないため距離の概念がない

KL情報量の差異をBoidの類似度と定義し，単語の連鎖の群れをつくる

- Boidの演算を導入することにより，単語の連鎖の群れに局所的に距離の概念が導入された。これにより単語の連鎖のBoidは半順序関係を持つ。
  よって、単語の連鎖の比較が可能となった。

比較可能にしていく過程

1. PMI　　前順序関係
2. Boid　　半順序関係
3. KL情報量　　完全律



単語の連鎖を集める(Boid:半順序関係)

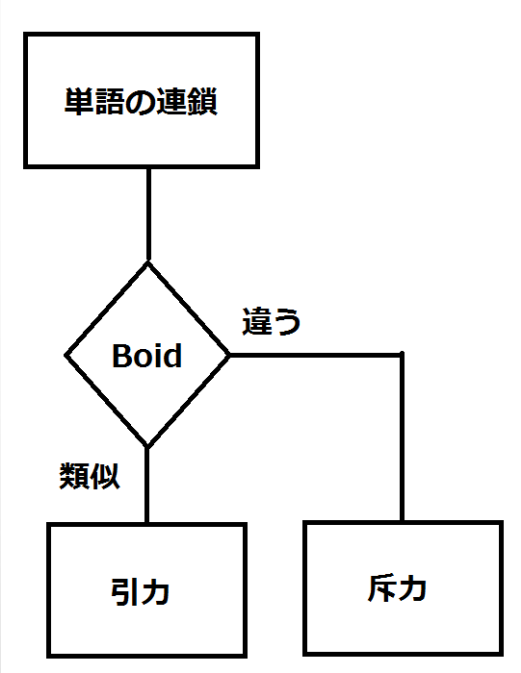
Boidとは
- 結合(Cohesion)

パーティクルが他の類似したパーティクルが集まっている群れの中心方向へ向かうように方向を変化させる
- 整列(Alignment)

パーティクルが他のパーティクルとだいたい同じ方向に飛ぶように合わせる
- 分離(Separation)

パーティクルが他のパーティクルとぶつからないように距離をとる

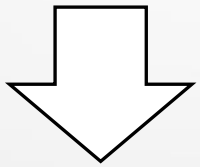
ここでは類似したパーティクル同士が群れの性質を満たすものとする



実際の利用方法

実験方法

- 確率的評価で集めた単語がテクストの文脈(潜在的意味)に関連することを試したい。これにより人間が実際に思う意味と情報量的評価との関連性を調べるためにこの実験を行う



単語の連鎖を実際に作成し、単語の連鎖によって単語が推定できるのかを実験によって確かめた.

実行環境

- OS:Ubuntu 16.04　windows10 1803(共に64bit)
- Anaconda 1.6.14(Python 3.6 or 2.7)…Python仮想環境
- CPU i7-7567U 3.50GHz
- メモリ 32GB(16GB×2)
- Gensim 3.4.0(word2vec 0.3.4)
- 英語wikipedia(15.5GB)

実験結果 1

数珠(数値は確率)

summer	→	team	→	she	→	final	→	third	→	she	ここで数珠が循環してしまった
	0.999849		0.999939		0.999949		0.999937		0.999945		

summerを含む文

1	At the 2016 Summer Olympics, she became the first female gymnast since Simona Am但nar in 2000 to win an all-around medal in two c
2	Her father, Farhat Mustafin, a Volga Tatar, was a bronze medalist in Greco-Roman wrestling at the 1976 Summer Olympics, and her moti
3	She competed twice over the summer, placing second in the all-around (58.250) at the Japan Cup in Tokyo in July and winning the all-a
4	At the end of July, Mustafina competed at the 2012 Summer Olympics in London.
5	In July, Mustafina competed at the 2013 Summer Universiade in Kazan, Russia, alongside teammates Nabieva, Ksenia Afanasyeva, Mar
6	At the 2016 Summer Olympics in Rio de Janeiro, Mustafina qualified to the all-around final with a total of 58.098, despite a fall on the ba
7	During the summer of 2001, students from Hillsboro's Miller Education Center worked at the park site to improve the riparian area of Ro
8	He won a gold medal at the 1924 Summer Olympics and a silver at the 1928 Games in the team sabre competitions.
9	He won a gold medal at the 1920 Summer Olympics and a bronze at the 1924 Summer Olympics.
10	He won a gold and bronze medal at the 1924 Summer Olympics.
11	He won the silver medal at the 2001 Summer Universiade.
12	His personal best times are 14:07.14 minutes in the 5000 metres, achieved in July 1996 in Hechtel; 29:38.88 minutes in the 10,000 metre
13	He won a gold and bronze medal at the 1924 Summer Olympics.
14	At 49 years and 141 days, competing on behalf of Israel at the 2004 Summer Olympics, he was the oldest track and field athlete compet
15	Running Movie (Original title in Hebrew: Seret Ratz), a documentary directed by Omer Peled and produced by Gidi Avivi in 2011, follows
16	Guido Balzarini (21 October 1874 – 1935) was an Italian fencer. He won a gold medal in the team sabre competition at the 1924 Summer
17	Valentino Argento (1901 – 8 September 1941) was an Italian fencer. He competed in the team foil competition at the 1924 Summer Olym
18	He competed at the 1920 and 1924 Summer Olympics.
19	He competed in the team foil competition at the 1924 Summer Olympics.
20	He won a gold medal in the team foil competition at the 1928 Summer Olympics.

下の文はsummerを含む文を検索した結果である。これをみるとこちらも数珠のそれぞれの単語が関係する文が表れている。

数珠

two	→	a	→	of	→	used	→	a
	0.999888		0.999958		0.999954		0.999955	

twoを含む文

At the 2016 Summer Olympics, she became the first female gymnast since Simona Am但nar in 2000 to win an all-around medal in two consecutive
She also qualified to the floor exercise final in third place, but withdrew and gave her spot to Grishina, who had been left out of the final due to the li
In qualifications, she fell on her first tumbling pass on floor (two whips into a double Arabian) and crashed her second vault (round-off, half-on, full
Hampered by an ankle injury, she performed on only two events in qualifications: uneven bars and balance beam.
In December, after competing for two seasons without a coach, she began working with Sergei Starkin, who coached world champion Denis Ablyazi
Two days later, Mustafina competed in the individual all-around final and scored 58.665 (15.200 on vault, 15.666 on uneven bars, 13.866 on balance
This made Mustafina the first female gymnast since Svetlana Khorkina to win the same event at two consecutive Olympics.
Two days later, after crashing her 1.5 Yurchenko and scoring a 12.433 on vault, 14.966 on bars, 12.533 on beam, and 13.066 on floor, she placed fourt
Myagdi Khola is a river which has its source at Mt. Dhaulagiri, then passes through Myagdi district to meet to the Kaligandaki river. The term "myagc
He was a member of their 1960 premiership team and two years later earned selection in South Australian state side but had to sit out due to a susp
He however performed well for St Kilda over the next two years.
During his last two years he was at the core, along with Richard Eden and George Batchelor, of founding the new Department of Applied Mathemati
The fortress is rectangular in shape, with two towers.
Beashel competed at the Olympics in the two-person keelboat, with Richard Coxon in 1984, Gregory Torpy in 1988, and David Giles from 1992 to 200
She and Adam have two sons, born in 2005 and 2008.
There they met their trainer, Kazbek, who had to choose the two best dogs from the group.
Chapter One (§ §1–4) covers the relationship between Norway and Svalbard; Chapter Two (§ §5–13) pertains to governance and courts; Chapter T
In analysis such as computational fluid dynamics (CFD), nanofluids can be assumed to be single phase fluids; however, almost all new academic p
An alternative approach simulates nanofluids using a two-component model.
Molybdenum disulfide (MoS2) and graphene work as third body lubricants, essentially becoming tiny microscopic ball bearings, which reduce the f
The band was formed by two teenage friends, bass guitarist Nikola Pavković and drummer Vladimir Jovanović, and named after a children's book.
The band did not record any material and after the two joined Instant Karma, the band ceased to exist.
The band's two songs, "Mladiću moj" and "Saša Ajdanov", appeared as soundtrack for the Darko Bajić movie "Balkanska pravila".
Pavković formed the band Kineska Kreda, and Dragana Mrkajić started working in the Zeleno Zvono club. Soon, the two returned to the original idea
While there, they are enticed by two prostitutes to join them at a private party way off the Strip.
Victor later wakes up in a cell in an abandoned building, and watches as two guards drag Anka out of her cell.
The four go to a nightclub, where they meet Kendra and Nikki, two escorts Carter secretly paid to have sex with Scott.
Two guards strap him to a chair in an empty room, with one wall made of glass, and Mike is on display to be gambled upon by wealthy clients.

こちらはあえてその単語自体にまったく他の意味を含まない「2」を選んだ結果である。この結果からは確かに単語同士が同じ文章内でよく使われていることがわかる。しかし、抽象的な単語ばかりになってしまい、実験結果として意味が薄いものができてしまったと思う。

- この実験では英語Wikipediaをword2vecによって単語の連鎖を作成したものである．上図はその一例．

実験結果２

- この実験では前回の結果をもとに簡単のために(実在は名詞に現れる)名詞のみを抜き出してword2vecを実行し単語の連鎖を作成したものである
- 名詞の抜き出しにはpolyglotを用いて行った

```
# -*- coding:utf-8 -*-

from polyglot.text import Text

t = "I eat an apple."
tokens = Text(t)
for token in tokens.pos_tags:
    print(token)
```

('I', 'PRON')
('eat', 'VERB')
('an', 'DET')
('apple', 'NOUN')

1.I　2.apple

名詞のみを抜き出したBoW

word2vec

単語の連鎖

Polyglotの使用例

pie	→	cake	→	dessert	→	pastry	→	dessert	
	0.801		0.802		0.8578				
					循環したので数珠を切った				
cakeを含む文									
Since sponge cakes are not leavened with yeast, they are popular ***dessert*** cho									
The cake was first invented by the Italian ***pastry*** chef Giovan Battista Cabona									

先ほどの実験と同じようにPMIの高いものを次の単語として選択しできた
単語の連鎖。連鎖としては意味の近いものが次の単語として選択されており、
連鎖として充分推定できると判断した.
これからわかることは名詞のみにすることで先ほどのような
連鎖としてふさわしくないデータが出にくくなった.

考察

Boidは似たもの(情報が似ているもの）が互いに引き合い，情報が似ていないものは互いに反発しあう性質を持つ．この性質そのものは半順序関係となる.

この性質を利用し，情報の近しいものを集める.

これにより情報セキュリティを分析させるための機械学習入力データを効率よく集めることが可能になる.

この結果を入力とする機械学習を用いれば,

情報漏えいを平均情報量によって評価可能となる見通しが得られる.