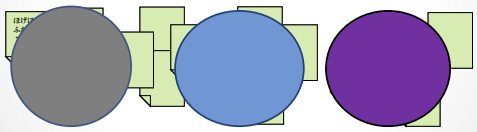
ニュースストリームの動的クラスタリング

広島大学大学院
情報工学専攻
小島　寛樹

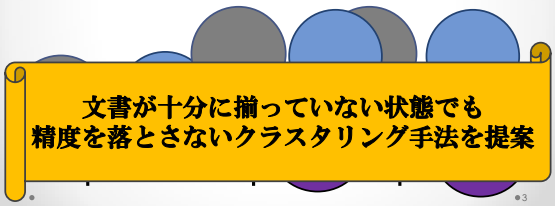
研究背景

- 一般的なクラスタリングでは文書集合は静的
  - 文書がすべて揃った状態で文書の重み付け・クラスタリングを行う



研究背景

- ストリームのクラスタリングでは文書集合は動的
  - ニュース記事は送られてきた時点でクラスタリングをしたい
  - 文書が揃っていない状態で文書の重み付け・クラスタリングを行う
- **静的なものと比べクラスタリング精度が低下**



文書が十分に揃っていない状態でも
精度を落とさないクラスタリング手法を提案

文書の重み付け

- 一般的な文書の重み付け手法：tf-idf
  - tf(term frequency): 単語の出現頻度
  - idf(Inverse document frequency): 逆文書頻度

$$tf_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,i}}, idf_i = log\frac{|D|}{|\{d: d \ni t_i\}|}$$

$n_{i,j}$は文書d_jに単語t_iが出てきた回数
$|D|$は総文書数
$|\{d: d \ni t_i\}|$は単語t_iを含む文書数

- ストリームの場合ではidfは動的に変化

idfの計算方法

- 一般的なidf

$$idf_i = log\frac{|D|}{|\{d: d \ni t_i\}|}$$

$|D|$は総文書数
$|\{d: d \ni t_i\}|$は単語t_iを含む文書数

- 文書が出現した日xまでの情報で計算するidf

$$idf_{i,x} = log\frac{\sum_{a=1}^{x}|D_a|}{\sum_{a=1}^{x}|\{d_a: d_a \ni t_i\}|}$$

分子は初日からx日までの総文書数
分母は単語t_iを含む初日からx日までの文書数

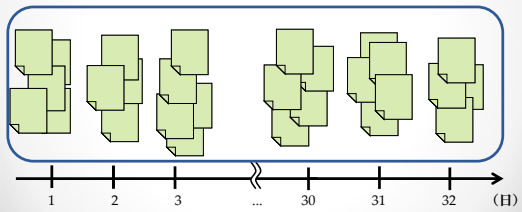
- 文書が出現した日xから過去1か月の情報で計算するidf

$$idf_{i,x} = log\frac{\sum_{a=x-30}^{x}|D_a|}{\sum_{a=x-30}^{x}|\{d_a: d_a \ni t_i\}|}$$

分子はx日から過去1ヶ月の総文書数
分母は単語t_iを含むx日から過去1ヶ月の文書数

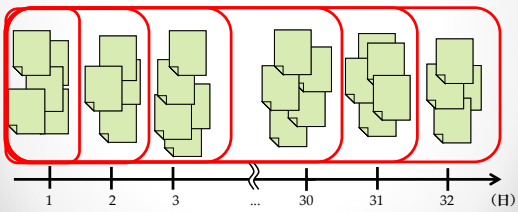
idfの計算範囲(1/3)

一般的なidfの計算範囲



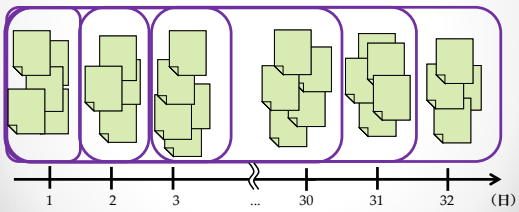
idfの計算範囲(2/3)

文書が出現した日xまでの情報で計算するidfの計算範囲



idfの計算範囲(3/3)

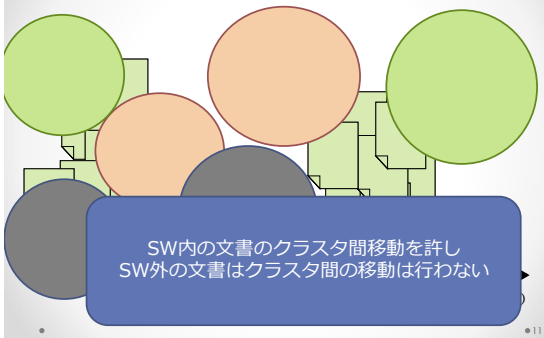
文書が出現した日xから過去1か月の情報で計算するidfの計算範囲



クラスタリング手法

- 一般的なクラスタリング手法:Kmeans法
  - Kmeans法ではクラスタ数は固定
- Kmeans法のアルゴリズム
  1. ランダムにクラスタを割り当てる
  2. 各クラスタの重心を計算
  3. 文書を最も近い重心を持つクラスタに割り当てる
  4. 割り当てに変化がなくなるまで2～3を繰り返す
- クラスタリングを行う範囲（スライディングウィンドウ:SW）を設定

クラスタリングの動き



実験

- Kmeans法でクラスタリングを行い、SW、idfの違いでの結果を比較
- 読売新聞のニュース記事
  - 2013年6月から2014年3月までの10ヶ月の記事
  - 記事数：35559
  - 特徴数：8340(MeCabを用いて形態素解析後、前処理を行った)
- SWのサイズ
  - 1週間(7日間)
  - 1ヶ月(30日間)
  - 6ヵ月(180日間)
- idfの計算方法
  - 全体で計算したもの(全体)
  - その日までの情報で計算したもの(その日まで)
  - その日から1ヶ月前までの情報で計算したもの(過去1ヶ月)

比較方法

- idfを全体で計算したもので重み付けをし、全体を見てクラスタリングをした結果とを以下の指標で違いを比較
  - purity(クラスタの純度)
  - cluster entropy(クラスタのエントロピー)
  - class entropy(クラスのエントロピー)
  - F-measure(F値)
    - エントロピーは値が小さいほど答えと近く、純度とF値は大きいほど答えと近い
- クラスタの中身を人の目で見て比較

結果(1/2)

- 全体を見てクラスタリングをした結果との比較

SW	idfの計算範囲	Purity	Cluster entropy	Class entropy	F-measure
	全体	21.4%	76.0%	78.4%	19.7%
1週間					
1ヶ月					
6ヶ月					

その日までの情報でのidfでは精度がより低下

過去1ヶ月の情報でのidfでは精度が低下しづらい

結果(2/2)

- クラスタの中身を人の目で確認

A　原子力規制委員会は、8日の新規制基準施行に伴い、電力会社が申請した原子力発電所の安全審査について…(13/07/09)

B　日本原燃は7日、青森県六ヶ所村にある使用済み核燃料再処理工場など、核燃料サイクル関連の4施設の安全審査を原子力規制委員会に… (14/01/07)

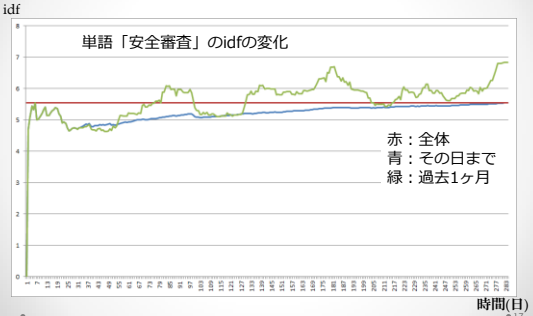
C　中部電力は6日、浜岡原子力発電所4号機（静岡県御前崎市）を再稼働させるための前提となる安全審査を…(14/02/06)

- これらの記事A,B,Cは1つのクラスタにまとめられるべき
  - SWが6ヵ月のものではまとめられている
  - SWが1ヶ月のものでは記事B,Cはまとめられている
  - SWが1週間のものではどれもまとめられていない
- 過去1ヶ月の情報でのidfを用いた場合、SWが1週間でも全てまとめられている

考察(重み付けについて)

- なぜ過去1ヶ月の情報での重み付けではクラスタリング精度が向上したのか
  - 過去1ヶ月に限定してidfを計算するということは全体をみると稀な単語でもある期間においては稀ではないこと、またその逆を反映できる
- idfの各計算方法でのある単語のidfの変化を調べ、クラスタリング結果への影響を分析

考察(重み付けについて)



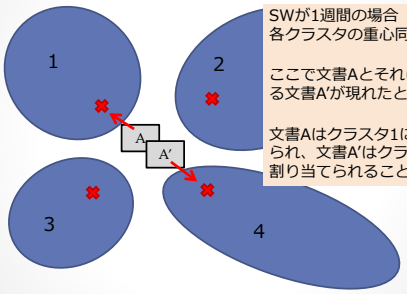
考察(クラスタリング手法について)

- SWが小さいとうまくまとめられていない
- コサイン類似度を用いて、各スライディングウィンドウでの各クラスタ間の類似度を計算
  - 類似度は0から1の範囲で表し、1が最も類似している

	1週間	1ヶ月	6ヶ月
類似度(平均)	0.355	0.117	0.074

- スライディングウィンドウが小さいほど各クラスタ同士は似ている

考察(クラスタリング手法について)

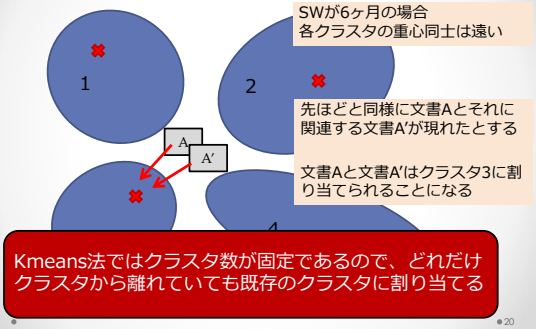


SWが1週間の場合
各クラスタの重心同士が近い

ここで文書Aとそれに関連する文書A'が現れたとする

文書Aはクラスタ1に割り当てられ、文書A'はクラスタ4に割り当てられることになる

考察(クラスタリング手法について)



SWが6ヶ月の場合
各クラスタの重心同士は遠い

先ほどと同様に文書Aとそれに関連する文書A'が現れたとする

文書Aと文書A'はクラスタ3に割り当てられることになる

Kmeans法ではクラスタ数が固定であるので、どれだけクラスタから離れていても既存のクラスタに割り当てる

まとめ

- ストリームのクラスタリングにおいて、SWが小さい場合、精度が落ちる
  - 各クラスタが近くにでき、うまく分類できない
  - トピック数が固定で無理矢理に文書を割り当ててしまう
- SWが小さい場合、文書の重み付けにおいて範囲を限定して計算を行うことで精度が落ちにくくなる
  - 単語の特徴をうまく拾い上げることができる