Simultaneous Clustering and Dynamic Keyword Weighting for Text Documents

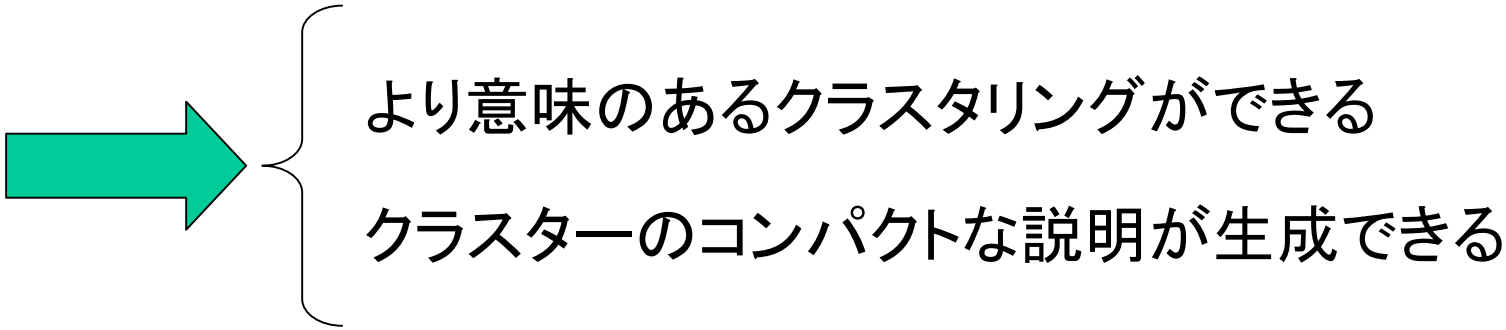
by

Hichem Frigui and Olfa Nasraoui

新納浩幸

Overview

クラスタリングを行うと同時に、クラスター内のキーワードの重みを同時に学習する手法（SKWIC）を提案

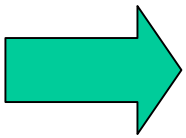


提案手法を拡張して soft クラスタリングもできるようにした

SKWIC

Simultaneous KeyWord Identification and Clustering of text documents

提案手法、k-means の拡張
各クラスタの term の重みを同時に計算



K-means で使う objective function に追加の項を加えただけ

K-means の利点がそのまま生かせている

Dissimilarity（1）

クラスター毎に term の
重みがあるのがポイント

i 番目のクラスタと文書ベクトル x_j 間の Dissimilarity

$$\tilde{D}_{ij} = \sum_{k=1}^{n} v_{ik} D_{ij}^{k}$$

n：文書ベクトルの次元数、つまり term の種類数

v_{ik}：i 番目のクラスタにおける k 番目の term の重み

$$V = \left[v_{ik} \right]$$

Dissimilarity（2）

文書とクラスターの
centroid との内積
つまり similarity

$$\tilde{D}_{ij} = \sum_{k=1}^{n} v_{ik} D_{ij}^{k}$$

$$D_{ij}^{k} = \frac{1}{n} - (x_{jk} \cdot c_{ik})$$

$x_j = (x_{j1}, x_{j2}, \cdots, x_{jk}, \cdots, x_{jn})$　文書ベクトル x_j

$c_i = (c_{i1}, c_{i2}, \cdots, c_{ik}, \cdots, c_{in})$　i 番目のクラスタ

Objective function

$$J(C,V)=\sum_{i=1}^{C}\sum_{x_i\in\chi_i}\sum_{k=1}^{n}v_{ik}D_{ij}^{k}+\sum_{i=1}^{C}\delta_i\sum_{k=1}^{n}v_{ik}^{2}$$

ただし　$v_{ik}\in[0,1]$　and　$\sum_{k=1}^{n}v_{ik}=1$

最小化

直感的には、クラスターを小さくしてゆくと第１項は小さくなるが、第２項は大きくなる。バランス。

Lagrange 乗数法の利用（1）

$$J(\Lambda,V)=\sum_{i=1}^{C}\sum_{x_i\in\chi_i}\sum_{k=1}^{n}v_{ik}D_{ij}^{k}+\sum_{i=1}^{C}\delta_i\sum_{k=1}^{n}v_{ik}^{2}-\sum_{i=1}^{C}\lambda_i\left(\sum_{k=1}^{n}v_{ik}-1\right)$$

V の行は独立（各 term の重みは独立）なので

各クラスタ i に関して

$$J_i(\lambda_i,V_i)=\sum_{x_i\in\chi_i}\sum_{k=1}^{n}v_{ik}D_{ij}^{k}+\delta_i\sum_{k=1}^{n}v_{ik}^{2}-\lambda_i\left(\sum_{k=1}^{n}v_{ik}-1\right)$$

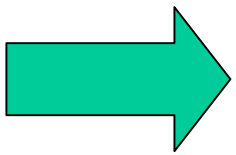
最小化

Lagrange 乗数法の利用（2）

目的関数 $$J_i(\lambda_i, V_i) = \sum_{x_i \in \chi_i} \sum_{k=1}^{n} v_{ik} D_{ij}^k + \delta_i \sum_{k=1}^{n} v_{ik}^2 - \lambda_i \left(\sum_{k=1}^{n} v_{ik} - 1 \right)$$

$$\frac{\partial J_i(\lambda_i, V_i)}{\partial \lambda_i} = \left(\sum_{k=1}^{n} v_{ik} - 1 \right) = 0$$

$$\frac{\partial J_i(\lambda_i, V_i)}{\partial v_{ik}} = \sum_{x_j \in \chi_i} D_{ij}^k + 2\delta_i v_{ik} - \lambda_i = 0$$



$$v_{ik} = \frac{1}{n} + \frac{1}{2\delta_i} \sum_{x_j \in \chi_i} \left[\frac{1}{n} \sum_{x_j \in \chi_i} D_{ij}^k - D_{ij}^k \right]$$

δ_i の設定（1）

$$v_{ik} = \frac{1}{n} + \frac{1}{2\delta_i} \sum_{x_j \in \chi_i} \left[\frac{1}{n} \sum_{x_j \in \chi_i} D_{ij}^k - D_{ij}^k \right]$$

δ_i が小さすぎると

1つのクラスタで1つの term だけに重み、あとは 0

δ_i が大きすぎると

全ての term に等しい重み

第1項と第2項を同じくらいの大きさになるように調整

δ_iの設定（2）

繰り返しの処理の中で設定してゆく

$$\delta_i^{(t)} = K_\delta \frac{\sum_{x_j \in \chi_i} \sum_{k=1}^{n} v_{ik}^{(t-1)} \left(D_{ij}^{k\,(t-1)} \right)}{\sum_{k=1}^{n} \left(v_{ik}^{(t-1)} \right)^2}$$

重みが負の場合

繰り返しの中では、重みがマイナスになる場合もある

頻繁に起こるようなら δ が小さすぎ

数回起こる場合は調整する

$$v_{ik} \leftarrow v_{ik} + \left| \min_{k=1}^{n} v_{ik} \right| \qquad \text{if} \quad v_{ik} < 0$$

初期の centroid は適当

初期のクラスタリング

最も近い centorid にデータを割り当てる

$$\chi_i = \left\{ x_j \mid \tilde{D}_{ij} \leq \tilde{D}_{kj}, \forall k \neq i \right\}$$

クラスター i とデータ j の dissimilarity

クラスター k とデータ j の dissimilarity

Centroid の更新

$$c_{ik} = \begin{cases} 0 & \text{if} \quad v_{ik} = 0 \\ \dfrac{\sum_{x_j \in \chi_i} x_{jk}}{|\chi_i|} & \text{if} \quad v_{ik} > 0 \end{cases}$$

クラスター i の中のデータ j の
次元 k（term k ）の値の和

SKWIC アルゴリズム（１）

初期設定

（１）クラス数 C を決める

（２）C 個の文書を適当に選び、それを centroid に設定

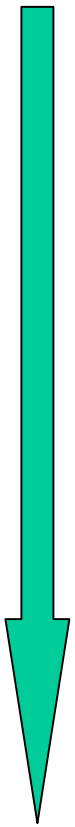
（３）最近傍法でクラスタリング

term の重みは全部 1/n

$$\chi_i = \left\{ x_j \mid \tilde{D}_{ij} \le \tilde{D}_{kj}, \forall k \ne i \right\}$$

SKWIC アルゴリズム（２）

繰り返し処理



（１）
$$D_{ij}^{k} = \frac{1}{n} - (x_{jk} \cdot c_{ik})$$

（２）
$$v_{ik} = \frac{1}{n} + \frac{1}{2\delta_i} \sum_{x_j \in \chi_i} \left[\frac{1}{n} \sum_{x_j \in \chi_i} D_{ij}^{k} - D_{ij}^{k} \right]$$

（３）
$$\tilde{D}_{ij} = \sum_{k=1}^{n} v_{ik} D_{ij}^{k}$$

SKWIC アルゴリズム（３）

繰り返し処理（続き）

（４）　再びクラスタリング

$$\chi_i = \left\{ x_j \mid \tilde{D}_{ij} \le \tilde{D}_{kj}, \forall k \ne i \right\}$$

（５）　Centroid を更新　$c_{ik} = \begin{cases} \dfrac{\sum_{x_j \in \chi_i}^{0} x_{jk}}{|\chi_i|} \end{cases}$

（６）　δ を更新　$\delta_i^{(t)} = K_\delta \dfrac{\sum_{x_j \in \chi_i} \sum_{k=1}^{n} v_{ik}^{(t-1)} \left(D_{ij}^{k^{(t-1)}} \right)}{\sum_{k=1}^{n} \left(v_{ik}^{(t-1)} \right)^2}$

以上の繰り返し

SKWIC vs. covariance matrix

SKWIC の term の重みづけ　≈　共分散行列の利用

違い）

共分散行列は term 間の関連性で重みをつけているのではない

➡ SKWIC はそう

共分散行列の利用にはデータにガウス分布の仮定がある

➡ SKWIC はそのような仮定がない

Soft/Fuzzy Clustering

Hard Clustering

データが1つのクラスに属する

Soft/Fuzzy Clustering

データが複数のクラスに属する、属する度合いがある

現実のクラスタリングでは Hard Clustering は難しい

Fuzzy SKWIC

Fuzzy partition

Fuzzy Clustering の出力

u_{ij} : データ x_j が i 番目のクラスタに属す度合い

$U = \left[u_{ij} \right]$ $C \times N$ の行列

$$\begin{cases} 0 \le u_{ij} \le 1 \\ 0 < \sum_{j=1}^{N} u_{ij} < 1 \\ \sum_{i=1}^{C} u_{ij} = 1 \end{cases}$$

Objective function

違いはここだけ

$$J(C,V)=\sum_{i=1}^{C}\sum_{x_i\in\chi_i}\sum_{k=1}^{n} v_{ik} D_{ij}^{k}+\sum_{i=1}^{C}\delta_i\sum_{k=1}^{n} v_{ik}^{2}$$

$$J(C,U,V)=\sum_{i=1}^{C}\sum_{x_i\in\chi_i}(u_{ij})^{m}\sum_{k=1}^{n} v_{ik} D_{ij}^{k}+\sum_{i=1}^{C}\delta_i\sum_{k=1}^{n} v_{ik}^{2}$$

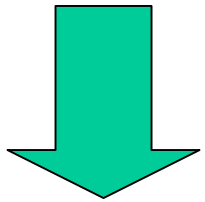
ただし　$v_{ik}\in[0,1]$　and　$\sum_{k=1}^{n} v_{ik}=1$

最小化

重みの式

Hard Clustering の場合と算出方法は全く同じ

$$v_{ik} = \frac{1}{n} + \frac{1}{2\delta_i} \sum_{x_j \in \chi_i} \left[\frac{1}{n} \sum_{x_j \in \chi_i} D_{ij}^k - D_{ij}^k \right]$$

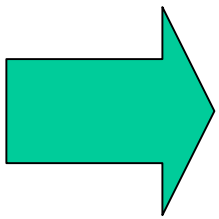


$$v_{ik} = \frac{1}{n} + \frac{1}{2\delta_i} \sum_{x_j \in \chi_i} (u_{ij})^m \left[\frac{1}{n} \sum_{x_j \in \chi_i} D_{ij}^k - D_{ij}^k \right]$$

δ_i の式

Hard Clustering の場合と算出方法は全く同じ

$$\delta_i^{(t)} = K_\delta \frac{\sum_{x_j \in \chi_i} \sum_{k=1}^{n} v_{ik}^{(t-1)} \left(D_{ij}^{k^{(t-1)}} \right)}{\sum_{k=1}^{n} \left(v_{ik}^{(t-1)} \right)^2}$$



$$\delta_i^{(t)} = K_\delta \frac{\sum_{x_j \in \chi_i} (u_{ij}^{(t-1)})^m \sum_{k=1}^{n} v_{ik}^{(t-1)} \left(D_{ij}^{k^{(t-1)}} \right)}{\sum_{k=1}^{n} \left(v_{ik}^{(t-1)} \right)^2}$$

Centroid の更新

Hard Clustering の場合と算出方法は全く同じ

$$c_{ik} = \begin{cases} 0 & \\ \dfrac{\sum_{x_j \in \chi_i} x_{jk}}{|\chi_i|} & \end{cases} \Rightarrow c_{ik} = \begin{cases} 0 & \text{if} \quad v_{ik} = 0 \\ \dfrac{\sum_{j=1}^{N} (u_{ij})^m x_{jk}}{\sum_{j=1}^{N} (u_{ij})^m} & \text{if} \quad v_{ik} > 0 \end{cases}$$

U の更新

$$u_{ij} = \frac{1}{\sum_{k=1}^{C} \left(\frac{\tilde{D}_{ij}}{\tilde{D}_{kj}} \right)^{1/(m-1)}}$$

Dissimilarity が負の場合

$$D_{ij}^{k} = \frac{1}{n} - (x_{jk} \cdot c_{ik})$$

これは負の場合もあり得る

$$u_{ij} = \frac{1}{\sum_{k=1}^{C} \left(\frac{\tilde{D}_{ij}}{\tilde{D}_{kj}} \right)^{1/(m-1)}}$$

これが負になるのはまずい

負の場合は調整

$$\tilde{D}_{ik} \leftarrow \tilde{D}_{ik} + \left| \min_{i=1}^{C} \tilde{D}_{ik} \right| \qquad \text{if} \quad \tilde{D}_{ik} < 0$$

Fuzzy SKWIC アルゴリズム（１）

初期設定

（１）クラス数 C を決める

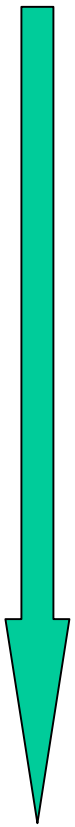
（２）m の値を適当に決める

（３）C 個の文書を適当に選び、それを centroid に設定

（４）U を適当に設定

Fuzzy SKWIC アルゴリズム（2）

繰り返し処理



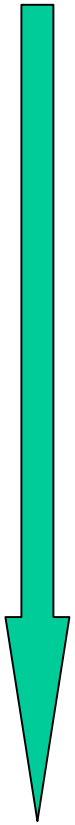
（1） $$D_{ij}^{k} = \frac{1}{n} - (x_{jk} \cdot c_{ik})$$

（2） $$v_{ik} = \frac{1}{n} + \frac{1}{2\delta_i} \sum_{x_j \in \chi_i} (u_{ij})^m \left[\frac{1}{n} \sum_{x_j \in \chi_i} D_{ij}^{k} - D_{ij}^{k} \right]$$

（3） 重みの調整

$$v_{ik} \leftarrow v_{ik} + \left| \min_{k=1}^{n} v_{ik} \right| \qquad \text{if} \qquad v_{ik} < 0$$

Fuzzy SKWIC アルゴリズム（３）



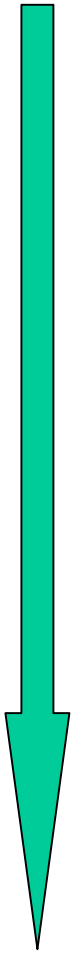
繰り返し処理（続き）

（４）
$$\tilde{D}_{ij} = \sum_{k=1}^{n} v_{ik} D_{ij}^{k}$$

（５）　$\tilde{D}_{ij}$　の調整

$$\tilde{D}_{ik} \leftarrow \tilde{D}_{ik} + \left| \min_{i=1}^{C} \tilde{D}_{ik} \right| \qquad \text{if} \quad \tilde{D}_{ik} < 0$$

Fuzzy SKWIC アルゴリズム（４）



繰り返し処理（続き）

（６）

$$u_{ij} = \frac{1}{\sum_{k=1}^{C} \left(\frac{\tilde{D}_{ij}}{\tilde{D}_{kj}} \right)^{1/(m-1)}}$$

（７）

$$c_{ik} = \begin{cases} 0 \\ \dfrac{\sum_{j=1}^{N} (u_{ij})^m x_{jk}}{\sum_{j=1}^{N} (u_{ij})^m} \end{cases}$$

Fuzzy SKWIC アルゴリズム（５）

繰り返し処理（続き）

（８）

$$\delta_i^{(t)} = K_\delta \frac{\sum_{x_j \in \chi_i} (u_{ij}^{(t-1)})^m \sum_{k=1}^{n} v_{ik}^{(t-1)} \left(D_{ij}^{k\,(t-1)} \right)}{\sum_{k=1}^{n} \left(v_{ik}^{(t-1)} \right)^2}$$

以上の繰り返し

実験）４クラス Hard Clustering

Web から news, business, entertainment, sports の文書
各50文書、計 200文書を収集、実験

Term の数は選別して 200、データは正規化

SKWIC は５回の繰り返しで収束

クラスタリング結果（Hard）

	Cluster 1	Cluster 2	Cluster 3	Cluster 4
	(business)	(entertainment)	(news)	(sports)
class 1	45	2	3	0
class 2	9	31	4	6
class 3	1	1	47	1
class 4	0	0	4	46

Table 3.1

高い重みの term（Hard）

ビジネス　　芸能　　ニュース　　スポーツ

Cluster # 1		Cluster # 2		Cluster # 3		Cluster # 4	
$v_{1(k)}$	$\omega_{(k)}$	$v_{2(k)}$	$\omega_{(k)}$	$v_{3(k)}$	$\omega_{(k)}$	$v_{4(k)}$	$\omega_{(k)}$
0.028	compani	0.031	fi lm	0.009	polic	0.021	game
0.015	percent	0.012	star	0.008	nation	0.013	season
0.010	share	0.010	dai	0.008	state	0.012	open
0.010	expect	0.010	week	0.008	offi ci	0.009	york
0.009	market	0.009	peopl	0.008	sai	0.008	hit
0.008	stock	0.008	like	0.007	kill	0.008	run

Table 3.2

実験）4クラス Fuzzy Clustering

m = 1.1

27回の繰り返しで収束

結論）

Hard のケースよりも若干精度がよい

クラスが誤ったデータは、他のクラスへの度合いも高い

クラスタリング結果（Fuzzy）

	Cluster # 1	Cluster # 2	Cluster # 3	Cluster # 4
	(business)	(entertainment)	(news)	(sports)
class 1	48	1	1	0
class 2	7	31	5	7
class 3	2	1	47	0
class 4	0	0	3	47

Table 3.3

ほとんど変化なし、若干よい

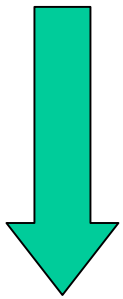
高い重みの term（Fuzzy）

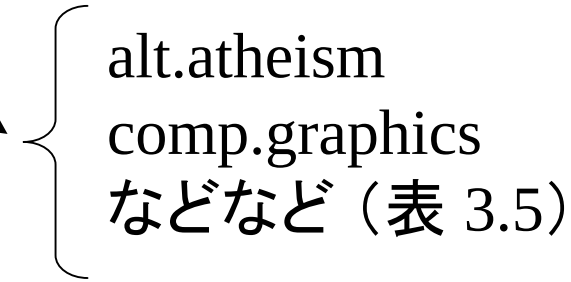
Cluster # 1		Cluster # 2		Cluster # 3		Cluster # 4	
$v_{1(k)}$	$w_{(k)}$	$v_{2(k)}$	$w_{(k)}$	$v_{3(k)}$	$w_{(k)}$	$v_{4(k)}$	$w_{(k)}$
0.029	compani	0.031	film	0.016	polic	0.025	game
0.016	percent	0.012	star	0.011	govern	0.015	season
0.011	share	0.010	week	0.010	state	0.010	plai
0.010	expect	0.008	dai	0.009	offici	0.009	york
0.008	market	0.008	peopl	0.009	nation	0.009	open
0.008	stock	0.008	open	0.009	sai	0.009	run

Table 3.4

実験）20クラス Hard Clustering

20 netnews group の各グループから 1000文書、
合計 20000文書





上記から 2000文書

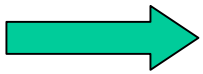
mini newsgroup data set

分析のために小さな
データも使った

実験は mini data と全体のデータの２種類行う

ミニ実験

mini newsgroup data set



前処理のために 1730文書に減少

利用した term は 449 種類

C＝40 で実験

本当は 20 クラス。細かく分けると、精度は高くなる。
クラスタリングの様子もよくわかる。

ミニ実験結果（1）

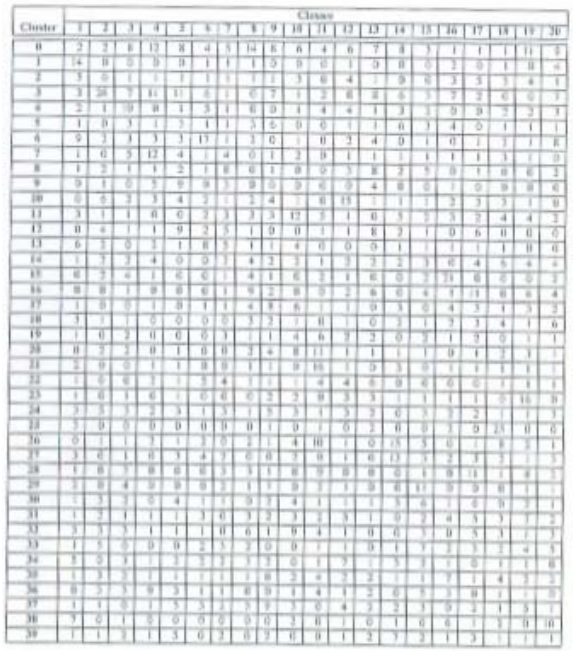
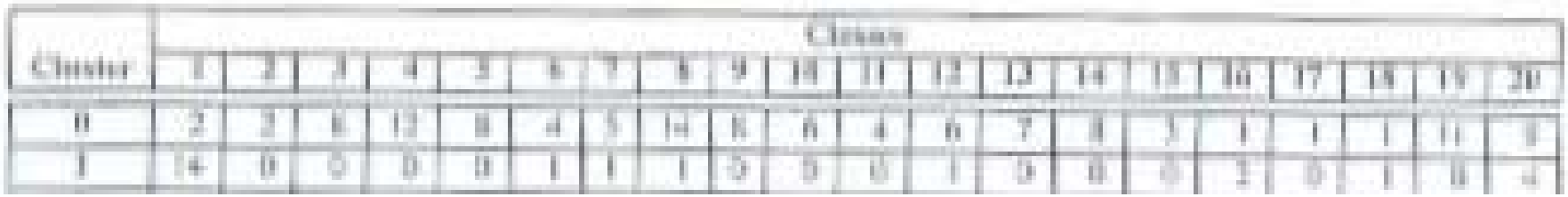


Table 3.6

結果は
めちゃくちゃ



ミニ実験結果（2）

Cluster	Card	Relevant Words
0	120	abort(0.1127); nord(0.0885); ec(0.0745); matt(0.0684); desktop(0.0538); coverag(0.0525); gordon(0.0554); backup(0.0476); er(0.0387); bill(0.0340);
1	25	atheism(0.8042); trap(0.0653); wisc(0.0228); smart(0.0157); protest(0.0071); dedic(0.0045); ownership(0.0038); absurd(0.0036); arriv(0.0036); probabl(0.0035);
2	34	senat(0.3504); dozen(0.2000); upset(0.1287); corrupt(0.1211); newspap(0.0164); motor(0.0161); remind(0.0076); loui(0.0067); weird(0.0062); point(0.0057);
3	108	cpu(0.1799); gui(0.1607); ei(0.0983); intel(0.0678); serial(0.0598); geneva(0.0541); ne(0.0317); adob(0.0421); shareware(0.0294); chi(0.0277);
4	42	appl(0.3982); mac(0.1896); haven(0.1554); ottawa(0.1136); sequenc(0.0179); invest(0.0096); survei(0.0068); forev(0.0053); ration(0.0051); visibl(0.0050)
5	37	nec(0.3993); babi(0.1853); johnson(0.1279); plate(0.0616); radio(0.0339);

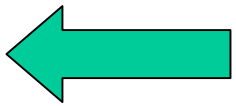
Table 3.7

ミニ実験結果、メモ

結果はメチャクチャに見える

重みがついた term はそれなりに納得できる

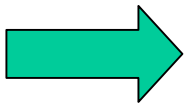
もともとクラスタリングが難しいデータ



Challenging Type のデータ
- noise データが多い
- クラスタが不明確
- ….

１６クラスへ

ある4つのクラスの文書がクラスタリングを難しくしている



これらの4つのクラスの文書を取り除いて、
１６クラスにして、再び、実験を行う

２０ クラスで行ったものと比較してみる

エントロピーによる評価

$$E = \sum_{i=1}^{C} \frac{N_i}{N} E_i$$

$$E_i = \frac{1}{\log K} \sum_{k=1}^{K} \frac{N_i^k}{N_i} \log \frac{N_i^k}{N_i}$$

C : 本当のクラス数

N : 全体の文書数

N_i : クラス i の文書数

K : 分類したクラス数

N_i^k : k 番目のクラスに分類されたクラスタ i の文書数

最終の実験結果

	Mini Newsgroup	Mini Newsgroup 16cls	20 newgroup
K Means	0.797	0.771	0.865
SKWIC	0.790	0.750	0.866
Fuzzy C Means	0.766	0.751	0.907
Fuzzy SKWIC	0.757	0.740	0.868

まとめ

- SKWIC : K-means の拡張手法、クラスター毎に term の重みを同時に学習
- Fuzzy SKWIC: SKWIC のソフト版、Fuzzy C-means よりも better
- Challenging type のデータに向いている
- K-means の拡張がそのまま使えそう、noise の扱い、クラスタ数の推定、大規模データの扱い、初期値の改良、、、など