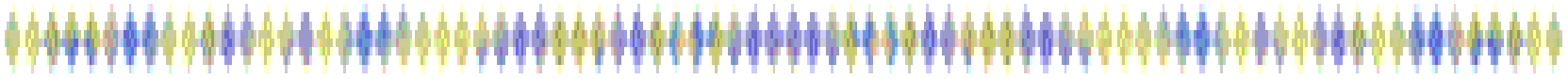
文書クラスタリングの技法ゼミ



C. 重みと類似度の計算方法

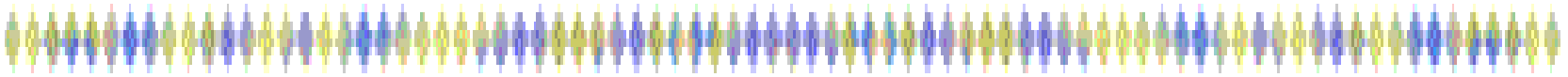
D. クラスタリングの結果の評価

E. 実験による性能比較

発表日：2007/6/13

発表者：阿部 竜之介

重みと類似度の計算方法



重み

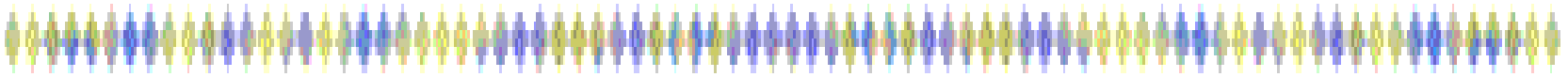
■情報検索理論（ベクトル空間、確率型など）に基づいて、複雑な式により計算する場合が多い

類似度

■重複係数やDice係数が使われることもあるが、余弦係数の適用例が多く、語の重みの計算に情報検索の理論を適用した場合には、その理論に沿ったかたちで計算がなされることもある

◆さまざまな重みの設定方法や類似度の計算方法のうちどれが優れているかという比較研究は少ない

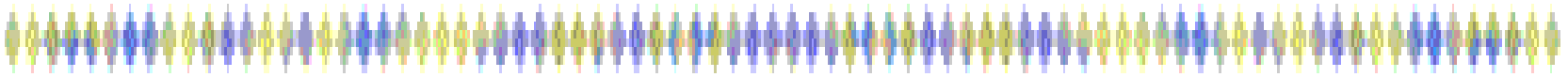
クラスタリングの結果の評価



クラスタリング結果の評価の類型

1. 直接的な評価
   （a）外部的な基準との評価（「正解」を使った評価）
   （b）内部的な評価
2. 間接的な評価

外部的な基準との評価



1.(a)外部的な基準との評価

各文書に分類記号が付与されているような「正解付き」の文書集合が利用できれば、クラスタリングの結果を直接的に、その「正解」という外部基準に照らして評価することが可能。

外部基準が利用可能なときの評価指標

1. F尺度
2. エントロピーまたは相互情報量
3. 正確性 …など

F尺度(1)

- F尺度は情報検索における伝統的な評価指標
- 再現率と精度との調和平均として計算される

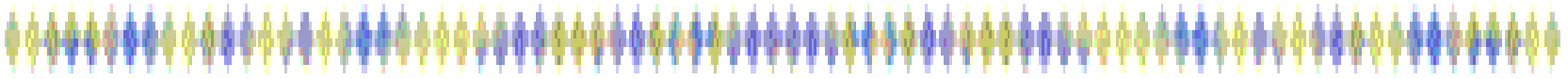
再現率 $R_e(K_h, C_k) = \frac{\tilde{n}_{hk}}{\tilde{n}_h}$

精度 $P_r(K_h, C_k) = \frac{\tilde{n}_{hk}}{\tilde{n}_k}$

「正解」にはH個の分類記号が含まれるとし、それぞれの分類記号が付与された文書集合を K_h と書く。

- 正解集合 K_h に含まれる文書数 $\tilde{n}_h$
- クラスタ C_k に含まれる文書数 $\tilde{n}_k$
- K_h と C_k とに共通の文書の数 $\tilde{n}_{hk}$

F尺度(2)



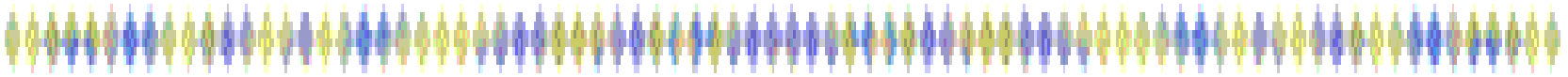
F尺度　$F_m(K_h, C_k) = \dfrac{2 \times R_e(K_h, C_k) \times P_r(K_h, C_k)}{R_e(K_h, C_k) + P_r(K_h, C_k)}$

この値は、ある特定の正解集合と特定クラスタの組ごとに計算されるので、クラスタリングの結果全体を評価するには、

$$F_s = \sum_{h=1}^{H} \frac{\tilde{n}_h}{N} \max F_m(K_h, C_k)$$

という指標が使用される。

エントロピー



正解集合 K_h にクラスタ C_k の文書が属する確率 $P(K_h \mid C_k)$ を考えれば、各クラスタのエントロピーを

$$E_k = -\sum_{h=1}^{H} P(K_h \mid C_k) \log P(K_h \mid C_k)$$

のように定義できる。

（ $P(K_h \mid C_k) = P_r(K_h, C_k)$ で推定）

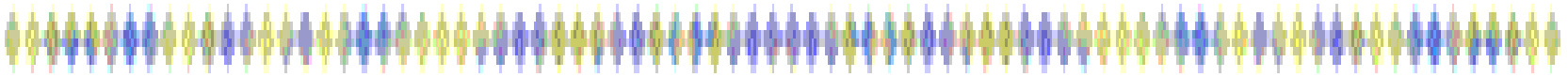
- E_k の値が小さいほど、望ましいクラスタリング

クラスタリングの結果全体に対しては

$$E(C \mid K) = \sum_{k=1}^{L} \frac{\tilde{n}_k}{N} E_k$$

で定義。

相互情報量(1)



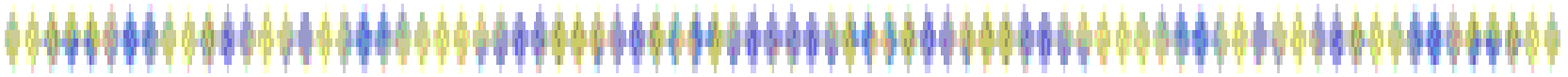
文書集合全体から1件の文書を選んだときに、それが正解集合 K_h とクラスタ C_k の両方に属する確率 $P(K_h, C_k)$ を考えて、相互情報量を

$$M_I(C,K) = \sum_{h=1}^{H}\sum_{k=1}^{L} P(K_h, C_k)\log\frac{P(K_h, C_k)}{P(K_h)P(C_k)}$$

で定義する。

- $P(K_h, C_k) = \tilde{n}_{hk} / N$
- $P(K_h) = \tilde{n}_h / N$
- $P(C_k) = \tilde{n}_k / N$　　で推定できる

相互情報量(2)



$M_I(C,K)$ の値は、正解集合とクラスタとが完全に独立ならば0、完全に一致すれば $\max(E(C),E(K))$ となる。

$E(C)$ はエントロピー

$$E(C) = -\sum_k P(C_k)\log P(C_k)$$

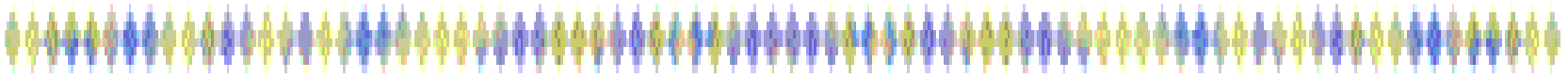
（$E(K)$ も同様）

最終的に

$$\hat{M}_I(C,K) = \frac{M_I(C,K)}{\max(E(C),E(K))}$$

とすれば、0から1の評価指標となる。

正確性



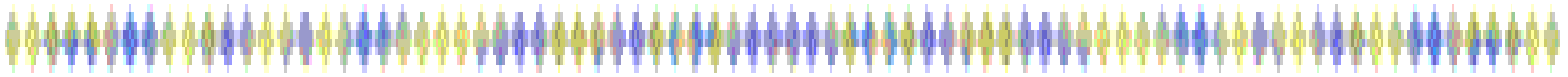
正確性とは

1つの文書に対して1つのクラスタが決定されたとき、もし、そのクラスタがその文書の属する正解集合に対応したものであるならば、「正しい」分類が行われたと判断できる。全文書中、何件が「正しく」分類されたかの割合を「正確性」と呼ぶ。

■正解集合とクラスタとの1対1の対応づけが容易ならば「正確性」を使うことができる

（ただし、正解集合 $K_1,\ldots,K_H$ も、クラスタ $C_1,\ldots,C_L$ もともに排他的で、 $H=L$ でなければならない）

内部的な評価

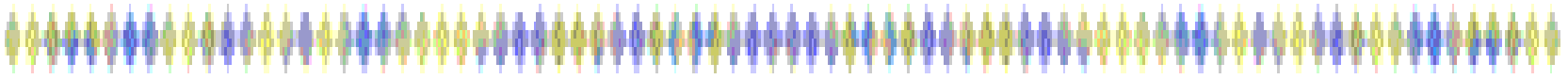


1.(b)内部的な評価

正解集合が利用できない場合に、クラスタ中の文書間に適当な類似度(または非類似度)を設定し、それに基づいて、クラスタがどれだけうまくまとまっているかを評価する。

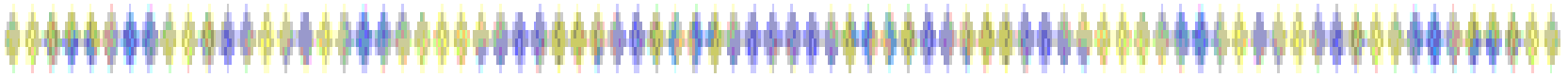
■1つのクラスタ中の文書相互間の類似度が高く、逆に他のクラスタに含まれる文書との類似度が低いほどクラスタリングに成功していると判断できる

階層的なクラスタリングの場合の評価



- 以上の各種の評価指標は非階層的なクラスタリングの場合にはそのまま使用できる
- 階層的なクラスタリングの場合には、樹形図の適当な結合レベルで階層を「輪切り」にし、クラスタの集合を設定する必要がある
- F尺度の場合にはあらゆる結合レベルでのクラスタでの最大値を使えばよいので、特に「輪切り」にする必要はない

間接的な評価

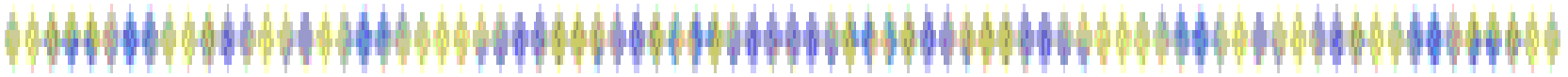


2.間接的な評価

例えば、文書クラスタリングの目的が情報検索であるときに、そのクラスタリングの結果を応用した情報検索の実行が、どの程度の検索性能の改善をもたらすかという点から評価するような場合である。

- この際には、検索実験用のテストコレクションをデータとして使用することができる
- 実際には情報検索の性能にはさまざまな要因が働くので、この方法によってクラスタリングの結果を独立的に評価することは難しい

実験による性能比較



- 文書集合の規模・性質に関する多様な条件の中で、クラスタリングの妥当性に関する客観的な知見を体系的に積み上げていくことは難しい
- ひとつの理想的な形は、TRECやNTCIRなどで試みられている、数多くの研究グループが参加する評価型ワークショップ（参加者同士の体験共有、意見表出、創造表現、意見集約を目的とした講座のようなもの）かもしれない
- 今後、実験による性能比較の結果を蓄積し、体系的な知見としていくことが必要