Recent Developents in Document Clustering

8 Phrase-Based Models

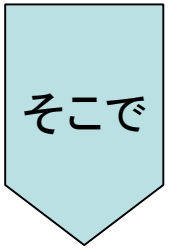
10NM733X　林華

8　フレーズベースモデル

フレーズ：単語の集まりで、文法構造が考慮されない

例：“the dog chased a cat”　と　“the cat chased a dog”

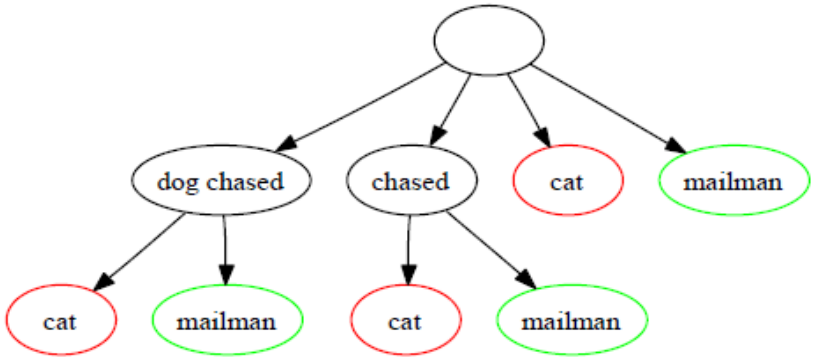
フレーズが同じだが、意味が違う。



仮説：語順を含んだフレーズのクラスタリングの精度が上がる

8.1　接尾辞木クラスタリング－１

- Suffix tree clustering (STC) :接尾辞木クラスタリングで、Webスニペットクラスタリングから由来



図：“the dog chased a cat”　と　“the cat chased a dog”の接尾辞木

- 利用法
  - コーパスの部分だけを使ってもよい ― 部分マッチ
  - クラスタリングに単語の頻度を用いる
  - 接尾辞がマッチするものだけを利用

8.1　接尾辞木クラスタリング－2

- 問題点
  - ―　大規模コーパスまたは長い文書の場合精度が低いあるいは高くない
- 解決策
  - ―　語順だけはクラスタリングの精度を上げることができないため、従来のベクトル空間モデルと併用する

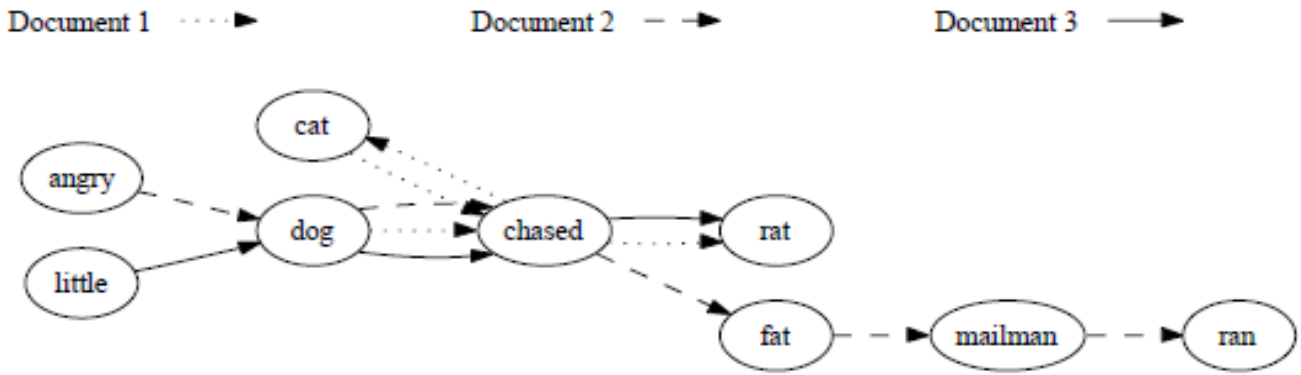
E＋, E－：文書の単語集

$$\varphi_{ST} = \frac{\left| E^{+} \cup E^{-} \right|}{\left| E^{+} \cap E^{-} \right|}$$

$$\varphi_{HYB} = \lambda \cdot \varphi_{ST} + (1 - \lambda) \cdot \varphi_{cos}$$

8.2　文書索引グラフ

- The Document Index Graph(DIG):文書索引グラフ
  例



- STCとDIGとの違いについて
  - DIGのほうが明確かつ単語頻度が記述されるので、余計な情報がない
  - DIGはクラスタリング手法ではなく、単なる記述法であり、他手法に使われる
  - マルチモデルの場合、その重みとしてDIGは70%～80%で、STCは30%～50%

■ 単語頻度プラス順序情報によるクラスタリングの精度が20%アップ