Recent Developments in Document Clustering

3. Vector Space Model

08T4083T 真下飛瑠

ベクトル空間モデル

- 元々は自動索引のために開発される
- m種類の単語が使われたn個の文書群をm×nの行列で表す
- 各文書はm次元のベクトル
- 重み付け方式では、単語の出現回数を利用している

tf-idf

- 最もよく使われる方式
- ある単語の文書内での出現頻度に、その単語が出てくる文書の、コーパス全体での出現頻度の逆数を掛けたものを重みに反映する
- コーパス全体での出現頻度は少ないが、その文書内で頻繁に出現する単語は重要、という仮定に基づく
- これらベクトルは正規化して行われるのが一般的

- ベクトル空間表現では単語の順序を考慮していない
- そのためbag of wordsやdictionaryモデルとも呼ばれる。
- 8章では語順情報を考慮した表現について述べる

２つの特徴

1. 一般的に、単語数が極めて大きくなり高次元の文書ベクトルになる（いわゆる“次元の呪い”）
2. コーパスはそれを構成する個々の文書が持つ単語よりもはるかに多い単語を持つため、ベクトルモデルでの行列がスパース行列になる

準備

原文をベクトルモデルにする前に、以下の処理が行われる

- Filtering
- Tokenization
- Stemming
- Stopword removal
- Pruning

- Filtering
  - ベクトルモデルでは意味のない文字や句読点を削除する
  - ウェブページのようなformatted documentsではより重要
- Tokenization
  - 文章を個々のトークン、言葉に分ける
  - 文法を解析し、品詞句や重要な単語を見分ける

- Stemming
  - 各単語を原型に戻す
  - “connected”, “connection”, “connections”は“connect”
  - Porter’s algorithmが一般的
- Stopword removal
  - Stopwordとはベクトル空間の次元で意味を成さないもの
  - 一般的な削除方法は、既知のStopwordと比較すること
  - 他に、品詞のタグ付けを行い、名詞、動詞、形容詞以外を削除する方法がある

- Pruning
  - コーパス全体で出現頻度の低い単語を削除する
  - 出現頻度の低い単語は役に立たない、という仮定に基づく
  - 場合により、出現頻度の高すぎる単語も削除することがある

例

- Document clustering has not been well received as an information retrieval tool. Objections to its use fall into two main categories: first, that clustering is too slow for large corpora (with running time often quadratic in the number of documents); and second, that clustering does not appreciably improve retrieval.

例

- document cluster receiv inform retriev tool object us fall main categori rst cluster slow larg corpora run time often quadrat number document second cluster appreci improv retriev

- 固有の単語は消さずに、全体の単語数は約半分になっている
- しかし中規模のコーパスでの文書ベクトルでは未だ数千の次元を持つ
- この問題の対処法は7章で議論する