トピックに基づく知識発見

2010年7月7日

知識工学部　阿部裕介

概要

1. トピックについて
2. 時系列ホットトピック抽出
3. ユーザ嗜好抽出
4. その他の応用

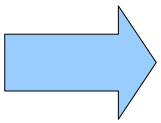
1. トピックについて

ベイズ的トピックモデルあるいはNMFなどによって得られたコーパス内の潜在的なトピックや特徴をここでは総称してトピックと呼ぶことにする

コーパス内の潜在的トピックが得られたとして、どのような応用が考えられるかについて紹介する

2. 時系列ホットトピック抽出

- 時系列情報を持つ文書群



月・季節・年などの単位でまとめてコーパス化

- このように構成されたコーパスのトピックを抽出することで、ホットトピックの時系列的変化を捉えることができると期待される

三月のコーパス

("Hot Topic "
#("梅" "ひなまつり"
"ホワイトデー"...))

四月のコーパス

("Hot Topic "
#("入学式" "卒業式" "桜"...))

五月のコーパス

("Hot Topic "
#("GW" "鯉のぼり" "メーデー"...))

トピック追跡モデル

- LDAのディリクレ事前分布に時系列順序の情報も加味した拡張版トピックモデルで、トピックの時系列変化を追跡する(Iwata et al, 2009)
- オンライン学習が可能で、累積していく

  大規模データにも対応できる

3. ユーザ嗜好抽出

- あるユーザの購買履歴や検索ログ等の総体をコーパスとし、そのトピックを抽出することで、そのユーザがどのような興味・嗜好を持っているかに関する知見を得ることが期待できる

- 似たような興味・嗜好を持つユーザをまとめてクラスタリングしたり、それを元に推薦システムを構築

- 十分なデータがあれば、あるユーザの興味・嗜好の時系列変化を追跡することも

- コーパスを取りまとめる視点を変えることで、異なる種類の知識発見が可能となる

- 時間 ⇒ トピックのトレンド変化
  人 ⇒ 各ユーザの興味・嗜好

これ以外の視点・切り口があれば、
また別種の知識発見も可能に！？

4. その他の応用

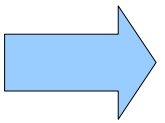
- 異常検出
- 画像認識
- ソーシャルアノテーション解析
- 文書可視化

異常検出

- トピック数の変化から新規トピックの出現を同定し、異常検出を行う(Yamanishi, 2010)

ソーシャルアノテーション解析

- ソーシャルアノテーションの例



はてなブックマーク
ニコニコ動画のコメント

ソーシャルアノテーションを分析して
Webページや動画に関する知見を得る
(Iwata et al, 2009)

画像認識

- 画像＝文書～Bag of Visual Words

  単語＝局所特徴量によるVisual Word

とみて、pLSAやLDAといったトピックモデルで解析する(Yanai, 2008)

文書可視化

- 従来の多次元尺度法に代わり、トピックモデルを用いて文書を可視化する手法(Iwata et al, 2009)

- 文書・トピックが二次元or三次元の可視化空間に座標を持つと仮定して生成モデルをつくり、潜在的トピックが近いものを近くに配置

参考文献

- 岩田具治：潜在トピックモデルを用いたデータマイニング, 2010
- 山西健司：潜在的構造変化検出の情報論的学習理論, 2010
- 柳井啓司：確率トピックモデルによるWeb画像の分類, 2008
- Iwata et al, Modeling Social Annotation Data with Content Relevance using a Topic Model, 2009
- Iwata et al, Probabilistic Latent Semantic Visualization:

  Topic Model for Visualizing Documents, 2008