

Application : 3

Fine-Grained Controversy Detection in Wikipedia

Siarhei Bykau, Flip Korn, Divesh
Srivastava, Yannis Velegarakis

名古屋大学情報科学研究科 近藤 弘隆

この論文について

- Wikipediaにおける議論の検知
 - 編集のポリシーの決定や必要な証拠がそろおう
 - 従来：ページ単位で破壊的編集を検知
 - より きめの細かい議論の発見が目的
- 議論とは
 - 同じ主題に対して複数人が関わった長期の論争



議論に関わる編集の集合を見つける

モデル

- Wikipediaの履歴：編集後の本文のみ
 - 単語に分割 or 記事中のリンクを利用
- Wikipediaにおける編集：追加，削除，置換
- 置換の主題：置換した単語と置換された単語のペア
- 同じ主題かどうかを比較したい
 - 編集した単語が同じでも同じ主題とは限らない
 - ✦ 例：“Polish” -> “French”
- 編集 e における半径 r の文脈 を定義
 - 編集によって置換された箇所の前 r 単語ずつ

議論の検知手法

- 入力：Wikipediaのページ編集履歴を使用
 1. 編集の抽出
 - 差分を利用して編集した箇所の特定
 - 置換した部分に注目
 2. 支持度の低い編集
 - 複数のユーザが編集していない編集を除外
 3. 文脈でクラスタリング
 - 編集 e における半径 r の文脈の類似度を利用
 4. 主題でクラスタリング，結合
 - 3のグループを主題で結合
 5. 議論の期間，関わった人の数，編集数で議論をランク付け
- 出力：議論の主題ごとに分かれた編集の集合

評価実験結果まとめ

- 再現率：Baselineをたいていの場合上回る
- 精度：ノイズは低く，適合率が高い
 - Baselineで見つけられなかった議論を多く発見
- テキストモデルの方がリンクモデルに対して多く議論を発見
 - しかし，ノイズも増える
 - ✦ 単語はリンクより更新されやすくノイズになりやすい
- 繰り返されない編集を除去することで大きくノイズを削減