大学共同利用機関法人 情報・システム研究機構
国立情報学研究所
National Institute of Informatics

トップエスイー ソフトウェア開発実践演習

FAQ作成支援のためのクラスタ内文書ランキング

株式会社富士通研究所　溝渕裕司　mizobuchi.yuji@jp.Fujitsu.com

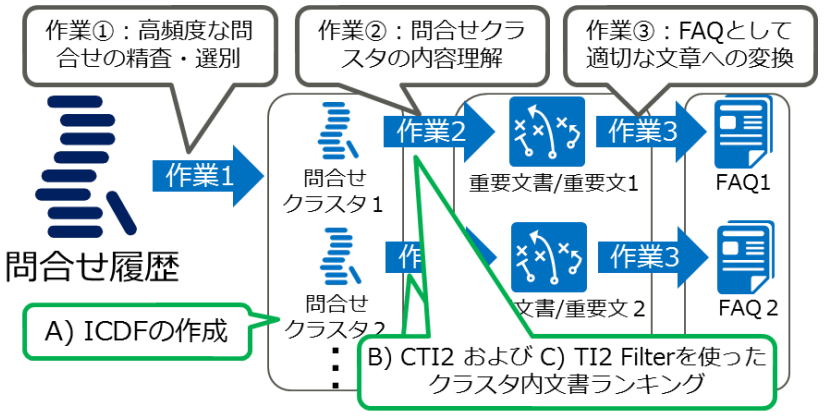
FAQ作成における問題点

- FAQ作成は過去の膨大な問合せ文書からの①精査選別、②内容理解、③文章作成を要し、多大な労力を要す
- なかでも内容理解は、膨大な文書読解が必要で的確な選別が必要である

解決方法

- クラスタ内文書ランキングの提案とその高精度化
  - A)：FAQ化タスクに特化した単語の重み付け方法の開発
  - B,C)：また、それを活用したクラスタ内ランキングの高精度化

FAQ作成の流れと提案手法



FAQ作成の流れと技術マッピング

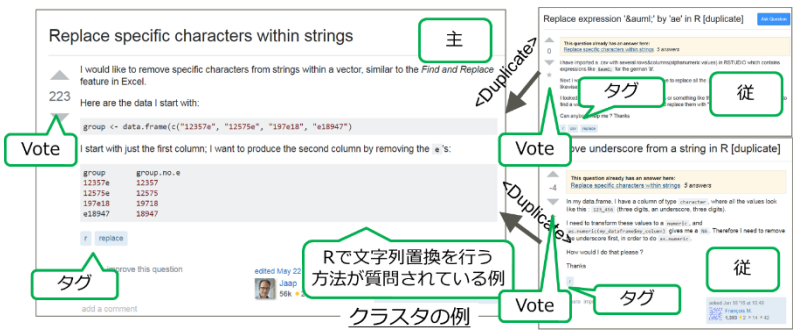
A) 新単語の重みづけ手法(ICDF)　新手法
- 文書全体では出現頻度が低く、同一クラスタ内では多くの文書で出現する単語を重要とする方法
- 通常のランキングで使われるTF-IDFをベースにICDF項を追加して高精度化を図る

B) 新文書メトリクス手法(CTI2)　応用1
- 文書内のICDF値の総和を文書の重要度とする方法

C) 新特徴量選択手法(TI2 Filter)　応用2
- ICDF値を使って単語の選択方法
- 今回は深層学習の判別モデルに適用

実験・評価方法

- データ：2018/9/5発行のStack Overflow Dump Dataを活用し以下を抽出
  1. 頻出する投稿のクラスタ
     - 3つ以上の文書からなる物で35352個抽出
  2. 各クラスタの正解文書順序をVoteから収集



- 評価指標：Adjusted-NDCG
  - 通常の文書ランキングの評価指標であるNDCG(Normalized Discounted Cumulative Gain)に補正項DCG_worstを加えたもの

結果

- CTI2の実験結果

機械学習モデル		メトリクス	A-NDCG	
教師なし		Tags Count	0.563	
		Cumulative TF-IDF	0.485	
		Gunning Fog Index	0.515	改善！
		CTI2	**0.563**	
教師あり	線形回帰	TM(Textual Metrics)	0.585	
		+RM(Readablity Metrics)		
		TM+RM+CTI2	**0.587**	改善！
	多層パーセプトロン	TM+RM	0.586	
		TM+RM+CTI2	**0.598**	改善！

- TI2 Filterの実験結果

Case	説明	入力サイズ（単語数）	隠れ層のサイズ	トータルパラメタ	A-NDCG
1	フィルタなし	2032504	20	40650142	0.688
2	**TF-IDF-ICDFによる単語選択**	**60549**	**50**	**3027602**	**0.701**
3	ランダムに単語選択	60549	50	3027602	0.554