特定分野における単語重要度計算手法Cr-Rvの提案と短い文章における 著者の専門性推定への適応

滝川 真弘,　山名 早人　(早稲田大学)

目的
短い文章から、著者の専門性を推定する

短い文章は情報量が少ない！ → 既存研究は… → 他の情報を用いて情報量を増やす → しかし… → ユーザによっては増やせない

SNS(Twitter)やQAサイト(Yahoo知恵袋)

過去の投稿[1]やサービス上のつながり[2]・貢献[3]

新規ユーザや非アクティブユーザ

golangのshared libraryとvim scriptの勉強として、簡単なスペル修正コマンドを作ってみた。本格的にやるなら辞書とかDBとか機械学習とかも取り入れたいな。

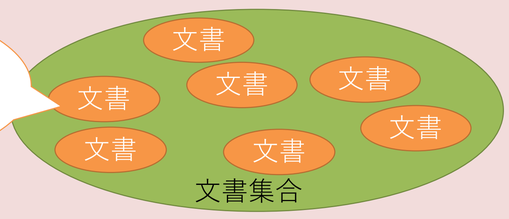
→ **専門性が高い (専門分野:プログラミング)**

どの単語が重要なのか → 重要度を付与する

関連研究

① **文書の中で**重要な単語
目的: 検索のためのインデックス付与

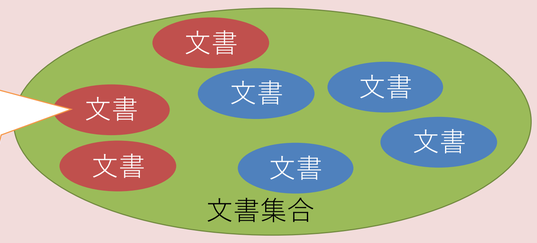
① 他の文書に対して、この文書を**特徴付ける単語**が重要



TF-IDF[4], BM25[5]

② **分野の中で**重要な単語
目的: 文書のカテゴライズ

② 他の分野に対してこの分野を**特徴付ける単語**が重要



tf*rf[6], tf*bdc[7], tf*PNF[8]

=> **目的が違う**

提案手法

- **分野性**
  - **一般分野の文書に出現する単語(専門用語)の重要度(専門度)は低い**
  - **一般分野に出現する単語の重要度を下げる。**
- **レア度**
  - **知名度が低い単語の重要度は高い**
  - **単純に低頻度の単語の重要度をあげるとノイズの重要度も上がる**
  - **各文書内における単語の出現頻度のエントロピーの逆数をとった**
- **ノイズの減少**
  - **どうしてもノイズ的な単語の重要度が上がってしまう**
  - **ノイズ的な単語は各文書内の出現頻度が低い**
  - **tfの最大値をかけることでノイズでない重要度をあげる**
  - **相対的にノイズの重要度を下げる**

$$\text{Cr-Rv(t)} = \text{Cr(t)} * \text{Rv(t)}$$

$$\text{Rv(t)} = \text{IH}(t) * \text{TFMAX}(t)$$

$$Cr(t) = \frac{\text{Norm}DF_P(t)}{NormDF_P(t) + \alpha * NormDF_N(t)}$$

$$\text{Norm}DF_P(t) = \frac{DF_P(t)}{|Dp|}$$

$$\text{Norm}DF_N(t) = \frac{DF_N(t)}{|Dn|}$$

$$\alpha = \frac{\sum_{t'}^{T} \text{Norm}DF_P(t)}{\sum_{t'}^{T} \text{Norm}DF_N(t)}$$

$$\text{IH(t)} = \log\left(\frac{\max_{t' \in T} H(t')}{H(t)}\right)$$

$$H(t) = -\sum_{d \in D} P(t,d) \log P(t,d)$$

$$P(t,d) = \frac{tf(t,d)}{\sum_{d'}^{D} tf(t,d')}$$

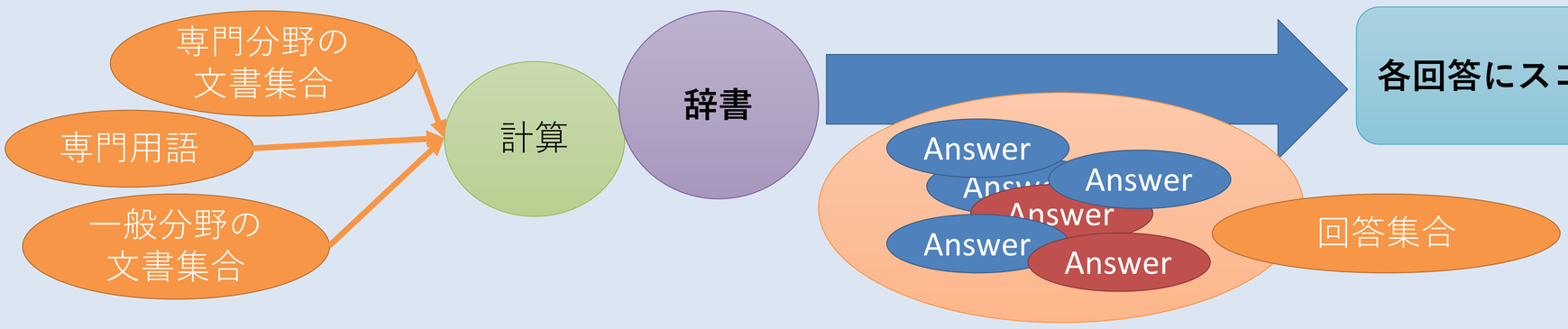
$$TFMAX(t) = \max_{dp \in DP} tf(t,dp) - \max_{dn \in DN} tf(t,dn) * \beta$$

$$\beta = \frac{\sum_{dp}^{DP} \sum_{t'}^{T} tf(t',dp) / |Dp|}{\sum_{dn}^{DN} \sum_{t'}^{T} tf(t',dn) / |Dn|}$$

t: 単語, |D|: 文書数, Dp: 専門文書集合, Dn: 一般文書集合, tf(t,d): 文書d内における単語tの出現頻度DF(t)$_P$: Dpにおける文書頻度, DF(t)$_N$: Dnにおける文書頻度,

実験

- ある回答の著者が専門家か否か(一般人か)
  - 専門家: **カテゴリマスター / 専門家/ その他専門性の高い職業**
  - 当てるのに何文字必要かも同時に調査



結果

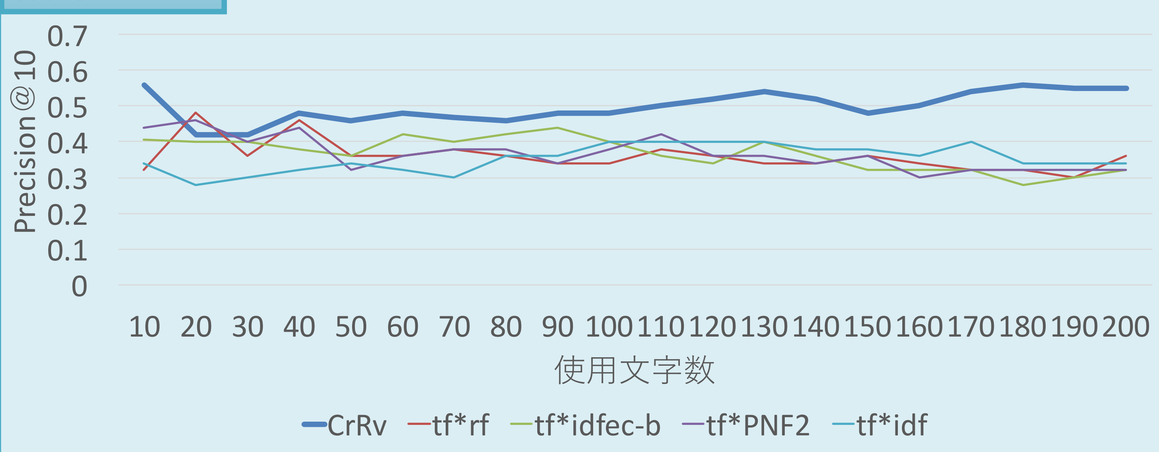


図 1 医療分野を特定分野とした時の各手法における
それぞれの文字数を用いた際のPrecision@10の値

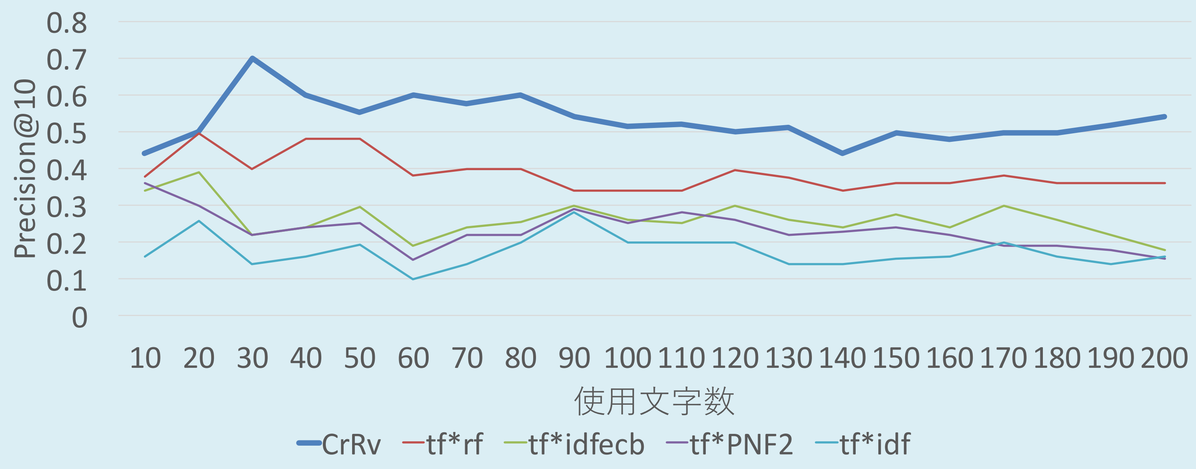


図 2 コンピュータ分野を特定分野とした時の各手法における
それぞれの文字数を用いた際のPrecision@10の値

まとめ

- 短い文書からの著者の専門性推定のため、
  新しい単語重要度計算手法 **Cr-Rv** を提案
- Yahoo知恵袋の回答から、著者が専門家か否かを推定
  医療 -> 最大56% / コンピュータ-> 最大70%
  今後の展望: 別カテゴリへの適用

[1] X.Shao, Z.Chunhong and J.Yang. "Finding Domain Experts in Microblogs" Proc. of the 10th Int'l Conf. on WEBIST (2014).
[2] Munger, Tyler, and Jiabin Zhao. "Identifying influential users in on-line support forums using topical expertise and social network analysis." *Advances in Social Networks Analysis and Mining (ASONAM), 2015 IEEE/ACM International Conference on*. IEEE, 2015.
[3] Lim, Wern Han, Mark James Carman, and Sze-Meng Jojo Wong. "Estimating Domain-Specific User Expertise for Answer Retrieval in Community Question-Answering Platforms." *Proceedings of the 21st Australasian Document Computing Symposium on ZZZ*. ACM, 2016.
[4] G.Saltion, E.A.Fox and H.Wu. "Extended Boolean Information Retrieval",CACM,Vol.26,No.11,pp.1022-1036 (1983).
[5] S.E.Robertson, S.Walker, S.Jones, M.M.Hancock- Beaulieu, and M.Gatford. "Okapi at TREC-3" , Proc. of TREC-3,pp.109-126 (1995).
[6] Deng, Zhi-Hong, et al. "A comparative study on feature weight in text categorization." APWeb. 2004.
[7] Domeniconi, Giacomo, et al. "A Study on Term Weighting for Text Categorization: A Novel Supervised Variant of tf. idf." DATA. 2015
[8] Naderalvojoud, Behzad, Ebru Akcapinar Sezer, and Alaettin Ucan. "Imbalanced text categorization based on positive and negative term weighting approach." International Conference on Text, Speech, and Dialogue. Springer, Cham, 2015.