概要

近年，一般ユーザが情報を発信する機会が増え続けている．そのような機会の中にコミュニティベースの Q&A (cQA) サイトが挙げられる．cQA はコミュニティのユーザが質問を投稿し，他のユーザが回答するサービスである．サービスの数やユーザが増していくに連れて，cQA サイトのアーカイブデータは豊富な情報源となっている．

この情報にアクセスするためには，Q&A アーカイブにおける検索が必要である．これはクエリ質問と類似した質問をアーカイブから検索することである．通常の文書検索と比較してのメリットとして，クエリが自然文なので情報要求をより的確に表現できること，関連する文書ではなく質問への回答を直接得られること，等がある．

類似質問検索には，文書が短いため word overlap が少ないという特有の難しさがある．これに対して取り組んだ既存手法では，複数の確率的モデルを手動で組み合わせたモデルを用いている．筆者はこの手法を拡張し，混合モデルとしての統一的なフレームワークを提案する．合わせて，混合モデルの構成要素として効果的なものの一例を提示する．

混合モデルでは混合比率の推定が必要となる．筆者は混合重みの扱いについて，Q&A によらず一定とするものと，Q&A ごとに異なるとする 2 つの方法を提案する．また，どちらの場合でも，訓練データからのモデルのサンプリング法およびその結果を用いた混合比率の推定方法を述べる．構成要素としては，既存手法において用いられたモデルをベースに考える．それらのモデルの表現する情報を考察し，質問文という言語現象を効果的に表現する構成要素を提案する．

これらの理論の有効性を検証するために実験を行う．まず，混合比率の推定について，効果的に推定するための学習データと基本モデルの組み合わせを模索する．次に，モデルに出現するパラメータの最適な値を，学習データサイズとの比較から検討する．最後に，上の 2 つの実験から得られた知見を基に，提案手法の性能を検証する．混合比率を一定とする手法と，Q&A ごとに異なるとする手法の性能を比較する．加えて，類似質問検索の既存手法とも性能を比較し，提案手法

の有効性が確認されたことを報告する.