システム情報工学研究科修士論文概要

年 度	平成 23 年度	学位名	修士(工学)
専 攻	知能機能システム 専攻	著者氏名	横本 大輔

指導教員氏名 宇津呂 武仁

論文題目

Wikipedia を知識源とする文書集合中の話題の同定と集約

論文概要

本論文では，ブログ空間における多種多様な話題を同定し，集約する方式について研究を行った．はじめに，Wikipediaを知識源として話題の体系を構築し．このWikipediaの体系を元に，ブロガーのブログ記事集合に対して話題を対応付ける方式を提案する．具体的には，Wikipedia中において初期クエリが出現するエントリを収集し，特定トピックにおける話題の候補とする．さらに，Wikipediaエントリ中の関連語の情報を利用して，ブログ記事を各話題に分類する．しかし，文書集合を効率よく俯瞰するためには，文書を話題に分類するだけではなく，複数の話題の間の類似関係を把握し，類似した冗長な話題を省き，代表的な話題に集約した上で閲覧する必要がある．この点において，上記の手法では，複数の話題の間の関連性を考慮した枠組みとなっていない．そこで，本論文ではさらに，複数の話題の間の冗長性を考慮して，文書集合における最適な話題俯瞰を実現するための文書クラスタリング手法を確立することを目的とする．

本論文の枠組みにおいては，まず，Wikipediaを知識源として，各文書の内容とWikipedia中の記述を照合しながら，各文書の内容に密接に関連した話題ラベルを複数個付与する．この話題ラベル付与の処理においては，各文書をクエリとして，Wikipediaのエントリを順位付けする問題として定式化し，特に，理論的な枠組みとして，クエリ尤度モデルに基づく方法を用いる．次に，話題ラベルが付与された文集集合に対して，複数の話題の間の冗長性を考慮して，文書集合における最適な話題俯瞰を実現するための文書クラスタリングを行う．以上の枠組みを，特定のクエリに対して関連するブログ記事を収集した文書集合に対して適用し，その有効性を評価した．本論文では，選定された上位のクラスタの正解率，および，上位クラスタ間の冗長性の除去に重点を置いてアルゴリズムの調整を行い，一部，既存研究を模倣した手法，および，語の頻度を用いたtf-idfによる手法等の性能を上回る性能を達成した．

審査日 平成 24 年 1 月 30 日

審査員	(大学名 職名)	(学位)	(氏名)
主査	筑波大学 准教授	博士(工学)	宇津呂 武仁
副査	筑波大学 教授	工学博士	白川 友紀
副査	筑波大学 准教授	博士(工学)	中内 靖