様　式　Ｃ－１９、Ｆ－１９、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

平成 ２８ 年　 ６ 月 １０ 日現在

機関番号：２４５０６
研究種目：若手研究(B)
研究期間：2012～2015
課題番号：２４７０００９７

研究課題名（和文）情報の詳細関係に基づくＷｅｂページの組織化

研究課題名（英文）Organizing Web Pages based on Detailing Relationship

研究代表者
湯本　高行（Yumoto, Takayuki）

兵庫県立大学・工学（系）研究科（研究院）・助教

研究者番号：２０４５３１５２

交付決定額（研究期間全体）：（直接経費）　　3,400,000円

研究成果の概要（和文）：ハウツー型情報，評判情報，ファクト型情報のそれぞれを対象に構成要素を抽出し，組織化して提示し，より詳細な情報へとナビゲートする手法を開発した．ハウツー型情報および評判型情報では機械学習を用いて，構成要素を抽出する手法を開発した．また，ファクト型情報ではニュース記事を対象として，類語辞書や編集距離などの文脈によらない手法と係り受け関係にある文節での対応関係などの文脈に依存した手法を組み合わせることで対応関係を発見する手法を開発した．

研究成果の概要（英文）：We focused on 3 types of information, how-to, review and news articles, and we proposed methods to extract elements from them. Then, we developed systems that organized them and navigated users to more detailed information. For how-to information and reviews, we use machine learning to extract the elements. For news articles, we find correspondence between the elements using context-independent methods and context-dependent methods. In the context-independent methods, we use synonym dictionaries and edit distance. In the context-dependent methods, we find on term correspondence in depending relationship.

研究分野：情報検索

キーワード：情報組織化　ハウツー情報　評判情報　機械学習

１．研究開始当初の背景

Web 上からユーザが望む情報をより見つけやすくするためにさまざまな検索技術が研究・開発されている．たとえば，Google で実現されているように，クエリ推薦によるユーザの検索意図の明確化などがある．これに関して，2006 年に実施された「情報爆発時代に向けた新しい基盤技術の研究プロジェクト」による 1000 人規模のアンケートである「情報検索に対する信頼性に関する調査」の中で調査が行われた．その結果によると，「Q3 検索を行うときの主な動機は何ですか」との質問に対し，80%以上が「検索キーワードについて今まで以上に深く知りたくなったため」と回答している．すなわち，詳細な情報に対する需要は非常に高いと言える．類似した研究として，用語の専門度や難易度の算出が提案されているが，内容の詳しさはこれらとは直交する概念である．一方，詳しさに基づく検索手法や詳しさを利用した検索システムは未だ実現されていない．

２．研究の目的

本研究では，情報の構成要素を抽出して，それらを組織化し，より詳細な情報へとナビゲートできるように提示する手法の開発を目的とする．対象とする情報のタイプは，ハウツー型情報，評判情報，ファクト型情報とし，それぞれに対して手法の開発を行う．

３．研究の方法

情報のタイプごとに以下の課題に取り組む．

(1)ハウツー型情報の組織化
(2)評判型情報の組織化
(3)ファクト型情報の比較提示

これらの課題では，情報の構成要素を抽出する手法を開発し，構成要素間の対応関係の発見手法についても開発する．その結果を用いて，情報を集約し，詳細な情報へとナビゲートするシステムの開発を行う．

また，上記の課題(1),(2)では機械学習を用いるが，大量の学習データを継続的に確保することも重要である．そのため，以下の課題にも取り組む．

(4)ソーシャルメディアデータを用いた擬似ラベル付けによる学習データの収集

この課題では，ソーシャルブックマーク(以下，SBM)のタグを用いて，擬似的にラベル付けし，それを学習データとして機械学習を行う手法を研究する．

４．研究成果

(1)ハウツー型情報の組織化

ハウツー情報は手順のリストとして表現し，手順は操作と対象のペアとして表現する．操作と対象の抽出方法はルールベースの方法と機械学習による方法の 2 つを検討した．

前者では，まず，過去形で書かれている文など明らかに手順を含まない文を文末表現に注目して除外する．次に，操作と対象を含む文節の直接的・間接的係り受け関係や含まれる助詞，助動詞などに注目したルールに基づき，操作および対象を抽出する．

後者では，各文節が操作および対象のそれぞれを含むかどうかを判定する分類を Support Vector Machine(以下，SVM)を用いて構築した．素性には対象の文節および係り先や係り元の文節に含まれる主辞の品詞や助詞などを用い，特定のジャンルに依存しにくい特徴量を使用した．これにより，後者の手法は前者よりも高い精度で手順を抽出することに成功した．

また，複数のハウツー情報掲載ページから手順の重要さおよび順序を決定することで要約を生成する手法について研究を行った．さらに，これらの手法を応用したハウツー情報の閲覧システムを開発した．このシステムでは，典型性によるフィルタリングや特定の手順を含むページの検索が可能であり，典型的な手順の把握や特徴的な手順の発見などが可能である．

また，重要性の新たな指標としてページと内容の含意関係の 2 部グラフにリンク解析を適用した手法を，独自性の指標としてコーパス中での出現確率を基にした非典型度の指標を開発した．

(2)評判型情報の組織化

評判情報については商品レビューを集約して提示するため，商品レビューから属性と意見のペアを抽出する手法を開発した．この手法では，係り先や係り元の品詞や文節に含まれる助詞など特定の商品カテゴリに依存しない特徴量を用いて素性ベクトルを構築し，属性および意見を含む文節かどうかを SVM でそれぞれ判定する．特に意見については，評価極性辞書を用いた方法に比べて，再現率を大きく向上させることができた．

さらに，抽出した属性-意見ペアを集約して提示するシステムを開発した(図 1)．このシステムでは，属性がタグクラウド形式(頻度の高いほど文字サイズが大きく表示される)で表示され，どのような属性があり，どれが重要かを直感的に理解できる．また，属性をクリックすると，意見がタグクラウド形式でポップアップし，これをさらにクリックすると，対応する文が表示されるため，詳細を知ることができる．

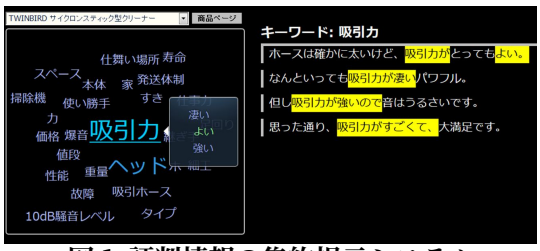

図 1 評判情報の集約提示システム

また，商品レビューには具体的なエピソードなど属性と意見を明らかに含まない文が

存在するが，そのような文をルールベースの方法を排除することで抽出精度が向上することがわかった．この適用範囲をさらに広げるため，前処理として，文の役割を判別する手法を開発した．この手法では文末表現などから素性ベクトルを構築し，SVM を用いて，文を意見，説明，エピソードに分類する．

(3)ファクト型情報の比較提示

ニュース記事を対象として，2 つの記事の間の内容的な共通部分を発見する手法を開発した．この手法では，文脈から独立した特徴を利用する方法と文脈に依存した特徴を利用する方法を組み合わせることで，表層的には違っていても意味的には共通する文を発見する．

前者では，語単位の対応関係を，編集距離，類語辞書，数値表現の場合は数値の近さを考慮して判定した．類語辞書は，Wikipedia のアンカーテキストとリンク先の関係から構築したもの，日本語 WordNet，含意関係辞書を組み合わせて用いることで網羅性の高い類語表現を獲得した．

後者では，対応関係にあると考えられる 2 文間で対応する文節の有無，それらの位置関係，文節間の係り受け関係を考慮することで，候補の選抜および追加を繰り返し行う．これにより，適合率，再現率の両面からの性能の向上を図った．

また，発見した共通部分以外の部分を差異とすることで，記事間の共通部分と差異を比較して提示するシステムを開発した(図 2)．このシステムでは，共通箇所が赤系，差異が青系で表現されており，どの部分が共通箇所または差異なのか，その割合はどのくらいかを知ることができる．

図 2 ニュース記事の共通部分・差異の比較システム

(4) ソーシャルメディアデータを用いた疑似ラベル付けによる学習データの収集

機械学習に用いる学習データの大規模化を容易にするために，既存のソーシャルメディアデータから擬似的な学習データを生成し，これを用いて分類器を構築する手法の開発を行った．特に SBM で学習し，ツイートの極性を分類する課題に取り組んだ．この課題では，Positive，Negative だけではなく，Positive を Interesting と Funny の 2 つに細分化し， Interesting，Funny，Negative とそれ以外に分類する．

まず，事前に SBM のタグと極性の対応関係について少量のルールを与えておき，それに基づき，SBM のコメントについて自動的にラベルづけを行った．このデータを用いて SVM で 2 値分類器を複数構築し，これを組み合わせて極性分類器を構築し，ツイートの極性分類を行った．

素性には，出現頻度が中程度の名詞を用いた．また，ツイートには口語表現やインターネットスラングなどが多く用いられる．このとき，複数の形態素解析器で解析すると結果が不一致になる場合が多いと考え，解析結果が一致するかどうかも素性として使用した．また，顔文字の有無も素性として使用した．

評価を行ったところ，素性を追加することによって分類精度が向上することが確認できた．一方，ツイートに人手でラベルづけしたデータを用いて学習した場合と比較したところ，提案手法は極性ごとに 300～350 件の学習データを用意した場合と同等の性能であった．

５．主な発表論文等

（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計 1 件）

① Akihide Bessho, Takayuki Yumoto, Manabu Nii, Kunihiro Sato, "Estimation of review helpfulness by content coverage and writing style", IADIS International Journal on WWW/Internet, 査読有, Vol.12, 2014, pp.102-114, http://www.iadisportal.org/ijwi/papers/2014121207.pdf

〔学会発表〕（計 24 件）

① Sho Iizuka, Takayuki Yumoto, Manabu Nii, Naotake Kamiura, "UHYG at the NTCIR-12 MobileClick Task: Link-based Ranking on iUnit-Page Bipartite Graph", NTCIR-12 Conference, 2016 年 6 月 9 日, 学術総合センター(東京都千代田区).

② 皿海宏明, 湯本高行, 新居学, 上浦尚武, "同じ出来事についての記事からの共通点と差異の抽出", 情報処理学会第 78 回全国大会, 2016 年 3 月 12 日, 慶應義塾大学矢上キャンパス(神奈川県横浜市).

③ 中野裕介, 湯本高行, 新居学, 上浦尚武, "機械学習による商品レビューの属性-意見ペアの抽出", 情報処理学会第 162 回データベースシステム研究発表会, 2015 年 11 月 26 日, 芝浦工業大学豊洲キャンパス(東京都江東区).

④ 山中隆広, 湯本高行, 新居学, 上浦尚武, "ページとクエリの連想確率に基づく希少な Web ページの検索", WebDB フォーラム 2015, 2015 年 11 月 24 日, 芝浦工業大学豊洲キャンパス(東京都江東区).

⑤ Yasuyuki Okamura, Takayuki Yumoto, Manabu Nii, Naotake Kamiura,

"Estimating Sentiment of Tweets by Learning Social Bookmark Data", 14th International Conference on WWW/Internet, 2015年10月24日, メイヌース(アイルランド).

⑥ 民岡佑規, 湯本高行, 新居学, 上浦尚武, "機械学習による商品レビューの文の役割の分類", 第14回情報科学技術フォーラム, 2015年9月17日, 愛媛大学城北キャンパス(愛媛県松山市).

⑦ 湯本高行, "機械学習によるハウツー情報の手順の抽出とその応用", 情報処理学会第77回全国大会, 2015年3月18日, 京都大学吉田キャンパス(京都府京都市).

６．研究組織

(1)研究代表者

湯本　高行（YUMOTO, Takayuki）
兵庫県立大学・大学院工学研究科・助教
研究者番号：20453152