様式 C-19

科学研究費補助金研究成果報告書

平成 23 年 5 月 31 日現在

機関番号：14301
研究種目：若手研究(B)
研究期間： 2008 ～ 2010
課題番号： 20700084
研究課題名（和文） 情報補完のための検索方式とそのクロスメディア検索への応用
研究課題名（英文）
Retrieval Methods for Information Complementation and Its Applications to Cross-media Information Retrieval
研究代表者
馬 強（Qiang Ma）
京都大学情報学研究科准教授
研究者番号： 30415856

研究成果の概要（和文）：
本研究では，情報の偏りを検知して補正できるシステム“情報栄養士”を実現するため，情報の相互関係を分析して組織化する技術，特に，話題分布および事象やエンティティの相互関係の分析とそれに基づく差異分析について研究開発を行った.また,応用システムを試作して，Web コンテンツ，放送コンテンツおよび新聞データを用いて実験を行い，開発技術の有効性と有用性を確かめた。

研究成果の概要（英文）：
In this work, to implement a information system called “Information Nutritionist”, which detects bias in information and provides complementary information to users, we are studying on relationship analysis and organization mechanisms of information; especially, we are working on analyzing topic distribution, causal and interests relations. Application systems based on these methods have been developed. We carried out experiments by using the data of Web contents, TV-programs and news articles. These experimental results have validated our methods.

交付決定額

（金額単位：円）

	直接経費	間接経費	合 計
2008 年度	1,400,000	420,000	1,820,000
2009 年度	1,100,000	330,000	1,430,000
2010 年度	800,000	240,000	1,040,000
総 計	3,300,000	990,000	4,290,000

研究分野：データベース、情報検索
科研費の分科・細目：情報学・メディア情報学・データベース
キーワード：情報補完、関係分析、マルチメディア、Stakeholder Model、発信者分析、因果関係、エンティティマイニング、バイアス

１．研究開始当初の背景

IT 技術の普及と進歩に伴って，発信され，再利用可能な情報量は，爆発的に増大している．爆発する大量で多様な情報からユーザの興味のある情報だけではなく，真に必要とする情報を効率よくかつ偏りなく獲得できることは，成熟したゆとりのある情報社会を実現するためには，必要かつ重要である．本研究では，情報の相互関係を整理して，偏りを発見して補正でき，バランスよく情報の獲得を支援するシステムを実現するための基盤技術について研究開発を行う．

２．研究の目的

本研究の目的は，情報の偏りを補正し，詳しくかつバランスよく情報を提供するシステム“情報栄養士”を実現するための基盤技術の確立である．そのため，以下の項目について研究開発を行う．

(1) Web ページの話題分布分析とそれに基づく差異分析の技術

(2) 事象間の関係分析と知識化技術

(3) エンティティマイニングおよびそれに基づく比較分析技術

３．研究の方法

本研究では，情報の偏りを明らかにするため，情報の関係を解明し，比較して差異を明らかする手法に着目して研究開発を行った．開発技術の有用性・有効性を保つため，要素技術とその応用の開発を同時に推進した．

(1) 話題と視点に基づく多様性分析技術：関連情報を収集して情報の分布を明らかにして差異を発見する手法を開発．

(2) エンティティマイニングとそれに基づく比較分析技術：エンティティに対する記述の差異を検知して発信者の偏りを検知する技術や，利害関係をマイニングして記述の差異を検知する技術を開発．

(3) 事象の因果関係マイニングと因果ネットワーク構築技術：事象の相互関係，由来を明らかにして知識化して情報の理解を支援する技術を開発．

４．研究成果

(1) 話題と視点に基づく多様性分析

インターネットで利用可能なニュースコンテンツも玉石混淆の状態になりつつある．特に，利用される素材や発信者の観点の違いから，ニュース内容の偏った場合がある．そこで，我々は，ニュース報道の多様性を分析し，偏りのあった場合はそれを明らかにしてユーザのニュース理解を支援するシステムについて研究開発を行っている．本研究では，我々は，記事や映像などのニュースコンテンツは実世界の部分クリピングであることに着目して，関連するニュースコンテンツをメディア横断して収集して分類と比較を行い，ニュース報道における視点（注目点）の多様性を分析してユーザに提示するシステムTVBanc（Topic and Viewpoint Based Bias Analysis of News Content）を開発した（図１）．

我々は，ニュース記事に書かれているイベントやアクティビリティを話題と呼び，それを表現するキーワードの構造体Content-Structure を提案する．Content-structure は，subject-term，aspect-term とstate-term から構成されるキーワードのタプルであり，語の頻度，品詞，位置および共起関係に基づいて抽出される．直感的に，subject-term は，イベントやアクティビリティの主題を表し，aspect-term とstate-term はそのアスペクトと状態を表す．本研究では，ニュースコンテンツのテキストの全体から抽出される content-structure をそのニュースの話題とし，ニュースの視点（注目点）を，追加説明や独自素材のよく書かれている最終段落から抽出されるcontent-structureとする．

与えられたニュース項目に対して，TVBanc は，まず，対象ニュースコンテンツの話題と視点を抽出して，関連ニュース報道をメディア横断して収集する．そして，これらのニュースコンテンツを分類して話題や視点の違いを分析し，エントロピーを計算して報道の多様性を評価し，結果をユーザに提示する．直感的に，多様性の大きいニュース項目の場合は，様々な視点の報道があるため，より多くのニュース項目を読んで理解する必要がある．一方，多様性の小さい項目の場合，その他の項目での新規情報が少ないため，多くのニュ

ース項目を読む必要はないが，情報の信憑性には疑問が残る．

10名の学生を被験者とし，NHK ニュース，FNNニュース，News-iとNews24から選んだ10個の映像ニュース項目を対象とした実験では，話題と視点の抽出手法および多様性分析手法の有効性と有用性の確認ができた．

図１ TVBanc

(2) エンティティマイニングとそれに基づく差異分析

①発信者分析

情報発信者が定常的に持っている独特の観点や取り上げる事象を分析して発信者のバイアスを明らかにすることが可能である．例えば，他の発信者に比べてある政党に対して否定的な記述が多い，ある国に関する記事ではある人物を一緒に取り上げがちであるといった特徴である．

我々はエンティティに関する記述に基づいてニュース発信者の特徴の差異を定量的に分析する手法を提案している．特定のエンティティに関する記述には発信者の観点や意見の違いが出ること多々ある．これは特定の政治家や政党，国に関する記述に関しては顕著であり，そういった記述は発信者の特徴を表しやすいものと考えられる．我々は各発信者の記事集合からこのような記述を分析し，発信者の特定エンティティに対する記述特徴を発見する手法を提案する．

ニュース発信者のエンティティに関する記述は次の2 通りに分けられると考えられる．

・エンティティに対する主観的記述

・エンティティに対する客観的記述

例えば「安部首相が悪い」といった記述は発信者の安部首相に対する否定的な意見が現れている．一方「安部首相が北朝鮮を非難する」といった記述は単なる事実を述べた記述である．我々は前者を発信者の主観的記述，後者を客観的記述と呼ぶ．主観的記述とは発信者の主観が現れているような記述であり，このような記述は発信者の特徴を直に表しているものと考えられる．客観的記述は事実を述べた記述である．客観的記述は単に事実を述べているだけであり，発信者の特徴には無関係であると考えられるかも知れないが，我々は発信者の事実の取り上げ方に特徴が表れていると想定している．例えばある発信者が同じ事実を何度もとりあげるといった特徴や他の発信者が取り上げていないような事実を取り上げているといった特徴である．このような事実の取り上げる回数や事実の網羅度を分析するために客観的記述は重要である．

特定のエンティティに関する記述を抽出するために，まず我々はそのエンティティに対して用いられている表現について分析する．エンティティに対して用いられている形容詞，副詞などは発信者の意見が現れていると考え，発信者の主観的記述に分類される．またエンティティが何をどうしたなど，エンティティが現れる「主語-目的語-述語」の表現は，エンティティに関する事実を述べていると考えられ，客観的記述に分類される．このようにエンティティに用いられる表現のパターンによって主観的，客観的記述を抽出する．注意したいのは，ここでは文単位ではなく，表現単位で記述を抽出する．

続いて抽出した主観的記述と客観的記述を

利用して，発信者の特徴ベクトルを生成する．特徴ベクトルを生成するにあたって，発信者の特徴を測るためのいくつかの指標を設定し，それぞれの指標に基づいて数値を算出し，ベクトルを生成する．まず一つ目の項目は，主観的記述に対する分析で，対象エンティティへの発信者の肯定，否定度を数値で表現する．次に二つ目の項目は，客観的記述に対する分析で，発信者が取り上げている対象エンティティに関する事実がバランスよいものか，偏ったものでないかを表す被覆度で数値化する．三つ目の項目は発信者が対象エンティティに対して主観的に書きやすいか，事実を取り上げる程度かを判定するためのもので，記述数で数値化する．これらの指標により発信者のエンティティに対する特徴ベクトルを生成する．生成した特徴ベクトルを用いて，発信者の過去の傾向の違いや発信者同士の傾向の違いを発見できることを，朝日新聞，毎日新聞，読売新聞および産経新聞のニュース記事（2010年1月1日から2010年8月20日まで）を用いた実験で確認した．

②ステークホルダマイニング

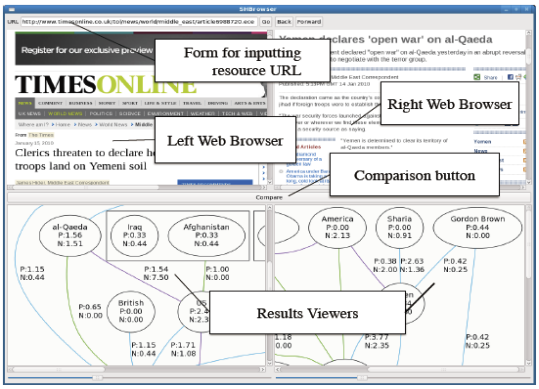

図２ ステークホルダマイニングによる比較

我々は利害関係者（ステークホルダー）に着目し，人物や利害関係に関する記事を分析して，ニュース報道における偏向の分析を行う手法を開発した．

我々は，ステークホルダーを「ある事象に参加しており，他の参加者とその事象に関して利害関係を持つもの」として定義する．また，利害関係を持つ可能性のある実体，例えば国，人，組織などをエンティティとする．ステークホルダーは他のステークホルダーと利害関係を持つため，利害関係を持つエンティティはステークホルダーと見なせる．よって，記事から事象における利害関係を分析することでステークホルダーを抽出できる．また，映像ニュースの場合，エンティティの出現頻度についても考慮する．

エンティティ間の利害関係の特定は文の構造を用いて行う．そこで我々はエンティティ間の関係を分析するために，関係の分析に適した関係構文構造を定義する．また，利害関係は単語によって表現されるため，エンティティの関係を表現する語彙の辞書である関係語辞書（RelationshipSentiWordNet）を構築した．この辞書は各語彙にポジティブ，ネガティブ，オブジェクティブの3 つの数値を割り当てており，ポジティブの数値が高い程エンティティ間の関係が良いことを表現し，ネガティブの数値が高い程悪いことを表す．オブジェクティブは客観的な程度を意味する．関係語辞書に含まれる語彙と関係構文構造を用いて記事から抽出した利害関係に基づいて利害関係グラフを構築する．このグラフから利害一致及び利害対立している関係をまとめることによって事象におけるステークホルダーを発見できる．そして，記事から抽出したステークホルダーと関係構文構造を用いて，記事のステークホルダーに対する評価を求める．評価値はWordNet から作成した感情辞書であるSentiWordNetを用いて分析を行う．

さらに，ステークホルダマイニングを用いた，ニュース報道の偏向分析手法を提案し，その試作システムを作成した（図２）．4人の大学生を対象としたユーザ実験では，ステー

クホルダマイニングによって事象に対するニュース記事の視点を定量化することができ，ニュース記事の偏向分析が可能となったことを確認できた．

(3) 因果関係マイニングと因果関係ネットワーク構築

① 因果関係の抽出手法

従来から因果関係の抽出や組織化に関する研究は数多く行われている．これらの研究では，言語的な特徴として手がかり表現を用いて因果関係を発見して抽出する手法を多く提案している．しかしながら，ニュース記事には，因果関係に関する記述の一部を省略する場合が多く，従来手法では因果関係の抽出を行うことが困難である場合がある．

そこで，本研究では，ニュース記事の主題と結果事象の一致性や，原因記述同士の類似性に着目し，因果関係を抽出する手法を開発した．具体的に，ナイーブベイズモデルを用いて記述の末尾の表現をモデル化して構築した識別器を用いて，タイトルと本文から結果と原因記述をそれぞれ抽出して因果関係を生成した．349件の経済ニュース記事を対象とした評価実験の結果から，記述が省略された場合でも，提案手法は，原因記述を3 割から4 割程度の再現率で抽出でき，さらに抽出された原因記述の約9 割は対応する結果事象と主題が一致していたことが分かった．因果関係抽出の適合率と再現率はそれぞれ0．74と0．65であった．

②因果関係ネットワークの増分構築手法

事象とその相互関係は，時間の経過とともに変化する．しかしながら，従来の因果関係ネットワークの構築手法では，因果関係の時系列特徴が十分に考慮されておらず，事象間の関係を正確に整理できない場合がある．

そこで，我々は，因果関係ネットワークを増分に構築する手法を提案する．基本的に，手がかり表現を用いて事象間の因果関係をニュース記事から抽出し，結合や簡略といった操作を用いて因果関係ネットワークを増分に更新していく．しかしながら，従来のキーワード集合ベースの事象の表現および結合手法では，因果関係ネットワークを効率よく構築できない．

・事象間の同一性判定を総当たりで行うため，計算量が大きい．

・事象の結合は事象を表すキーワード集合の類似度に依存するため，結合の順序によって生成されるネットワークが異なる場合がある．つまり，因果関係ネットワークの構築の一貫性に問題がある．

そこで我々は，キーワードの抽出手法に述語項構造を導入し，事象をトピックと内容を表すキーワード構造体のペアによって表現する TEC（Topic Event Casual network Model)モデルとそれに基づくネットワークの結合手法を提案し，これらの課題の改善を試みた．

・トピックと内容を独立して事象を表現するモデルを用いて事象間の同一性判定を類似トピック内に限定することによる計算量の削減を実現した．

・事象を表現する語彙の役割（SVO 構造）と意味（概念）を考慮してネットワークを構築する手法を提案し，事象の内容を表す語の役割を考慮し，WordNet やオントロジーを用いて概念レベルで事象の類似性を計算することで，ネットワーク構築の一貫性を保つことを可能となった．

我々は，GoogleNews から七つのトピック計60 記事を収集して評価実験を行い，ネットワーク構築の効率の改善を確認できた．

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２件）

(1) Tatsuya Ogawa, Qiang Ma, Masatoshi Yoshikawa, News Bias Analysis Based on Stakeholder Mining, IEICE TRANS. INF. & SYST., Vol. E94-D, No.3, pp 578-586, 2011, 査読有

(2) Shin Ishida, Qiang Ma, Masatoshi Yoshikawa, Extraction of Characteristic Description for Analyzing News Agencies, Journal of Digital Information Management, Vol.8, Issue 6, pp 349-355, 2010, 査読有

〔学会発表〕（計２４件）

(1) 小川達也，新規性の高いユーザ生成コンテンツの自動発見，第 19 回 WI2 研究会，2011 年 3 月 8 日，東京，査読無

(2) 野島裕輔，ニュース記事の内容と構造特徴を考慮した因果関係マイニング，DEIM2011，2011 年 2 月 27 日，静岡県，査読無

(3) 有光淳紀，ユーザー体験指向の Twitter 検索手法，DEIM2011，2011 年 2 月 27 日，静岡県，査読無

(4) 許レイ，Discovering Inconsistency in Multimedia News Based on a Material-Opinion Model，HICSS2011，2011 年 1 月 7 日，ハワイ（米国），査読有

(5) 石田晋，発信者のエンティティに対する記述特徴分析によるニュースの推薦，WebDB Forum 2010，2010 年 11 月 11 日，東京，査読有

(6) 石井裕志，概念と構造を考慮した事象の類似性判定に基づく因果関係ネットワークの増分構築，WebDB Forum 2010，2010 年 11 月 11 日，東京，査読有

(7) 許レイ，Exploring Special Items in Multimedia News Based on a Stakeholder Model，WI2010，2010 年 9 月 1 日，トロント（カナダ），査読有

(8) 小川達也，Stakeholder Mining and Its Application to News Comparison，WI 2010，2010 年 9 月 1 日，トロント（カナダ），査読有

(9) 石井裕志，An Incremental Method for Causal Network Construction，WAIM2010，2010 年 7 月 17 日，四川省（中国），査読有

(10) 石井裕志，因果関係ネットワークの増分的な構築について，第７２回情報処理学会全国大会，2010 年 3 月，東京，査読無

(11) 小川達也，ステークホルダマイニングとそれに基づくニュース報道の比較，DEIM2010，2010 年３月，兵庫県，査読無

(12) 石田晋，記述の主観性を考慮したニュース発信者の特徴分析とその応用，DEIM2010，2010 年３月，兵庫県，査読無

(13) 石井裕志，Casual Network Construction to Support News Understanding，HICSS2010，2010 年 1 月 6 日，ハワイ（米国），査読有

(14) 許レイ，Stakeholder Extraction for Inconsistency Analysis of Multimedia News，WebDB フォーラム 2009，2009 年 11 月 20 日，東京，査読有

(15) 石井裕志，SVO 構造を用いた因果関係ネットワーク構築手法について，DBS 研究会，2009 年 11 月 20 日，東京，査読無

(16) 石田晋，Analysis of News Agencies' Descriptive Feature Using SVO Structure，ICDIM2009，2009 年 11 月 4 日，アナーバー（米国），査読有

(17) 石田晋，Analysis of News Agencies' Descriptive Feature of People and Organization，DEXA2009，2009 年 9 月 3 日，リンツ（オーストリア），査読有

(18) 桐谷雄介，Classifying Web Pages by Using Knowledge Bases for Entity Retrieval，DEXA 2009，2009 年 9 月 3 日，リンツ（オーストリア），査読有

(19) 馬 強，Topic and Viewpoint Extraction for Diversity and Bias Analysis of News Contents，APWeb/WAIM2009，2009 年 4 月 2 日，蘇州（中国），査読有

(20) 石井裕志，因果関係ネットワークの構築によるニュースの理解支援，DEIM2009，2009 年 3 月 9 日，静岡県，査読無

(21) 桐谷雄介，エンティティ検索支援のための知識ベースを用いた Web ページ分類，DEIM2009，2009 年 3 月 9 日，静岡県，査読無

(22) 石田晋，ニュースサイトによる人物や組織に関する記述の特徴分析，DEIM2009，2009 年 3 月 9 日，静岡県，査読無

(23) 馬 強，話題と視点に基づくニュースコンテンツの多様性分析システム，第 13 回 WI2 研究会，2008 年 12 月 11 日，神奈川県，査読無

(24) 馬強，Ranking People Based on Metadata Analysis of Search Results，E-BAG 2008，2008 年 9 月 2 日，オークランド（ニュージーランド），査読有

〔その他〕

ホームページ等

http://www.db.soc.i.kyoto-u.ac.jp/soc/

６．研究組織

(1)研究代表者

馬 強　　（Qiang Ma　）

京都大学情報学研究科准教授

研究者番号：30415856