様式Ｃ－１９

科学研究費助成事業（科学研究費補助金）研究成果報告書

平成２４年　４月２７日現在

機関番号：14301
研究種目：基盤研究(B)
研究期間：2008～2011
課題番号：20300036
研究課題名（和文）　Web 上の知識資源の統合利用基盤技術に関する研究

研究課題名（英文）　Fundamental Technologies on Integrated Usage of Knowledge Resource on the Web

研究代表者
吉川　正俊（YOSHIKAWA MASATOSHI）
京都大学・大学院情報学研究科・教授
研究者番号：30182736

研究成果の概要（和文）：Web サーバにおいて高品質な情報を管理するために，情報の注釈データの管理手法を開発すると共に，構造化文書の照応解析技術を開発した．知識資源を表現する RDF データの格納及び検索システムを構築すると共に，検索エンジンと利用者生成型知識資源 Wikipedia の統合利用システムを開発した．また，複数ニュースソースデータの統合利用手法として整合性提示機能提供システムおよび因果関係ネットワーク漸増構築法を開発した．

研究成果の概要（英文）：We have studied management methods for annotation data on information, and coreference analysis in order to keep the quality of information on Web servers high. We have developed a system for storage and retrieval of RDF data representing knowledge resource as well as integrated usage system of retrieval engines and user generated knowledge resource Wikipedia. Also, we have developed a system for consistency analysis and a method for incremental update of causal relation networks.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
2008 年度	4,500,000	1,350,000	5,850,000
2009 年度	3,800,000	1,140,000	4,940,000
2010 年度	3,300,000	990,000	4,290,000
2011 年度	2,900,000	870,000	3,770,000
年度			
総　計	14,500,000	4,350,000	18,850,000

研究分野：　情報学
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：知識資源、Web, XML, マルチメディア，情報統合

1.　研究開始当初の背景

Web 上の情報資源を表現する表現モデルとして RDF などの提案はあったが，Web 上の知識資源を表現する RDF データに対する効率性の高い格納，検索システムは提案されていなかった．また，Web 上の一般的なデータと異なる性格を持つ利用者生成型知識資源が増大しつつあったが，それらを統合的，補完的に利用する手法や，複数のニュースソースなどの知識資源を統合利用する手法が十分には開発されていなかった．Web 上のデータの品質については，データの由来などを記述する注釈データの情報がデータ演算後も保持できるモデル，システムの提案はなかった．

２．研究の目的

本研究では，Web 上で公開され質の保証がさ

れている知識資源を統合的に利用することにより，爆発的に増大し続けている Web 上のデータからより品質の高い情報を利用者の嗜好に応じて容易に探索可能とするための基盤技術について以下の課題を中心に研究を行う．

(1) 情報を提供するサーバ側において情報の由来など信憑性の判断材料を提供する手法や秘匿すべきプライバシ情報を含むコンテンツを検出する技術．

(2) サーバ側やクライアント側において大量の RDF データを効率的に管理，検索する技術．

(3) 知識資源を利用した高品質な情報の探索およびその結果の効果的統合と意味的相互補完技術

３．研究の方法

(1) 大量 RDF データの格納及び検索システムの構築

RDF のためのハイブリッド型スキーマを用い関係スキーマの動的な変化を許容する格納方法を研究する．また，ある情報資源に関連性が高い範囲など RDF の問合せ言語である SPARQL の表現範囲を超える問合せを可能とする言語の開発を行う．RDF の直列化表現として利用される XML 文書についてもキーワード集合を対象とする意味的に妥当な検索結果を与える

(2) 情報の内容に基づくプライバシ情報の検出

膨大な利用者生成コンテンツの中からプライバシを侵害していると考えられるデータをサーバ側で検出する方法を開発する．秘匿すべき情報を単語列で与え，文中の照応関係やオントロジデータを利用することによりプライバシ情報を含む箇所を高精度で検出する技術の研究を行う．

(3) 情報の由来などのデータを提供する手法

サーバにおいて情報を保持する関係表に対する演算が実行された場合でも，情報の由来などに関する注釈を保持可能とするデータモデルを開発し，注釈を効率的に管理する手法を研究する．

(4) 検索エンジンと知識資源の統合利用

利用者が情報探索のために入力として与えた簡易自然言語問合せ文を構文解析し，それを基に検索エンジン，オントロジデータ，ドメインデータの結果を段階的に相互利用することにより，問合せに対して，知識ベースや検索エンジン単独では得ることができない高品質の結果を提示するための基礎研究を行う．

４．研究成果

(1) 情報の注釈データの管理手法

膨大な履歴情報の系統的な管理を支援するとともに，データ利用者に対して履歴情報の検索，閲覧を可能とする手段を提供する枠組みを開発した．多様な粒度のデータに対して付与された注釈を処理を通して伝播させる手法を提案し，実験により従来手法に対する空間コスト，時間コスト面での優位性を示した．

また，データ処理によって起こる注釈の品質劣化の問題を解決するため，注釈の意味を表現するための表現モデルを提案した．さらに，注釈の意味的な定義に従い，データ変遷において生じる注釈の不整合を定式化するとともに，注釈の整合性を維持するために，注釈付与の根拠となった情報を注釈と共に伝播させることによって，注釈の客観性を維持した注釈管理手法を提案した．注釈を生成する方法には，注釈生成者が問合せを発行し，その結果に対して注釈を関連付ける暗示的な方法と，注釈の付与対象となるレコードを生成者が個々に指定する明示的な手法がある．明示的な手法に対しては，生成者が付与対象を決定した根拠となるデータに関する客観的な情報が明らかでないため，そのままでは根拠に基づいた注釈管理手法を適応することができない．明示的な注釈の付加範囲を，等価な暗示的表現に変換する手法を示すし，特に問題となる，付与対象となるレコードの絶対数が少ない場合のために注釈の内容を考慮した分割評価指標を考案することによって，注釈生成者の意図に近い条件を優先的に選択することを可能とし，より正確に注釈の付加条件を推定する手法を与えた．従来の決定木推定手法との比較実験により，すべての場合において提案手法が優位であり，特に，課題であった学習データが十分に用意できない場合においても従来手法に対する優位性が見られた．

(2) 構造化文書における照応解析

Web 上で頻出する構造化文書におけるプライバシ情報検出のための基礎技術として，文書の論理的構造の位置関係から求められる照応の起きる確率の利用を構造による照応とし，自然言語的な素性に加えて構造化文書における文書構造の素性を用いた照応解析を行った．文書構造の特徴として，構造化文書から，照応木パターンと呼ばれる照応の出現関係を抽出した木構造を用いる．照応木パターンをもとに，いくつかの文書構造の素性を提案し，構造化文書を入力として，自然言語的な素性と文書構造の素性を抽出し，それをもとに機械学習により照応関係を解析した．Wikipedia を XML 化した文書集合をベ

ンチマークとして，文書構造の素性による照応解析の実験を行い，訓練データとテストデータが同文書から生成される場合，自然言語的な素性のみを用いた従来の照応解析のF値，提案手法である自然言語的な素性と文書構造の素性を統合した照応解析のF値はそれぞれ 68.5%, 77.1% となり，文書構造の素性を用いることによって，評価が向上し，構造化文書において本手法が効果的であることを確認した．

(3) 大量 RDF データの格納及び検索システムの構築

RDF における特定のノードに関する知識資源を表現するサブグラフとして既に提案されている Concise Bounded Description(CBD)を基に，述語に重み付けをすることにより一般化を行った Dynamic Concise Bounded Description (DCBD)および RDF のための問合せ言語 DCBDQuery を提案した．DCBDQuery は，DCBD を構築し DCBD に対して意味を持つ経路および最短経路を見つけるために用いられる．さらに，RDF データを関係データベースに格納するための Updated Schema-aware 表現を考案した．RDF グラフは内部のリンク文から主記憶内で生成される．また，データベースと主記憶内グラフのハイブリッドデータモデルにアクセスするための DCBDQuery 問合せエンジンを設計した．DBLP++の RDF データを用い，既存の代表的な RDF データベース Jena2 との比較を行うことにより，提案アプローチが効率的であることを示した．

直列化された RDF の高速検索に関しては，RDF を直列化する際に用いられる XML のキーワード検索では，最小共通祖先(LCA)を求めることが一般的であるが，ID/IDREF により異なる木構造間に概念的なつながりがある場合は LCA 全体またはその部分を返すことは不適切である．そこで概念的なつながりを捉えるために Smallest Lowest Object Tree(SLOT)および Smallest Interrelated Object Tree (SIOT)を提案し，これらを用いることにより意味的な結果を返せることを確認した．また，利用者支援の観点から，問合せに対する検索結果と問合せの関連語との関係性など，初期問合せを別のどのような表現で代替または近似代替できるかを提示するための代替表現の表現法および代替表現獲得の高速化法を開発した．

(4) 検索エンジンと知識資源の統合利用

エンティティ検索のための知識ベースを利用した Web ページの分類に関し，知識資源を利用した高品質な情報の探索およびその結果の効果的統合と意味的相互補完技術として，Web 検索時の目的の一つであるエンティティ検索のために知識ベースを用いることにより検索 Web ページを分類する新たな手法を提案した．本手法は，Web ページと Wikipedia 記事の類似度を計算し，YAGO を参照しながら Web ページ，エンティティ，カテゴリの間の有向グラフである PFC グラフを構築する.PFC グラフを参照することにより Web ページをカテゴリに分類する．実験により本手法の有効性を確認した．

Wikipedia のような Web 上の利用者生成型事典を利用した情報検索に関する研究として，Wikipedia の文書間リンク構造を減衰流を用いて解析することにより，概念や事物間の関連の強度を求める手法を開発した．また，始点，終点間の経路に出現する関連を裏付ける情報を利用することにより，二つの事項に関連のある画像を精度良く検索するための手法を開発した．さらに，Wikipedia の記事とその記事における画像のキャプションの整合性，および記事タイトル，画像キャプションに対する画像の典型性，代表性を計算する手法を開発した．また，Wikipedia の編集履歴情報を利用し編集者の信頼性を計算することにより Wikipedia 記事の信頼性を推定する手法を開発し，その手法に基づき実際に提示されている Wikipedia 記事における各部分文書の信頼度を利用者に提示するシステムを開発した．

また，Wikipedia の記事とそれに関連した Web 文書を取得し，それらを LDA を用いて解析することにより，Wikipedia の内容を補完するトピックに関する Web 文書を高い確率で取得する手法を開発し，記事に含まれていない新たな情報を検索する能力である精度，検索された文がどの程度良くグループ化されたものであるかを示す純度，本手法により選択された潜在トピックが人間によって評価されたトピックをどの程度包含するかを示すトピック包含度の三つの尺度において，提案手法の優位性を示した．

(5) 複数ニュースソースデータの統合利用

複数のニュースソースデータによる人物や組織などのエンティティに対する記述の主語，動詞，目的語の三つ組みを比較，解析し，新たに構築した感情辞書を利用することにより各ニュースエージェンシによるエンティティに対する記述極性を抽出し，信憑性の基となる情報の整合性を提示する機能を提供するシステムを開発した．また，各ニュースエージェンシのニュースを時系列的に解析することにより通常とは異なる特異な論調のニュースを検索する手法を開発した．

また，ニュースイベント間の関連を明確化するために因果関係の漸増的構築方法を提案した．因果ネットワークモデルとしてトピック/イベント因果(TEC)モデルを用い，それ

に基づく漸増的構築方法を与えた．因果関係の連鎖を得るためには同じイベントを表すノードを検出する必要がある．トピックキーワードが概念レベルで表されるトピックに限定するにより検出時間を短縮化する．キーワードの意味同定の方法を提案するとともに，キーワード比較の三種類の意味距離を導入することにより，我々の方法により従来手法に比べ同様のイベントを表すノードをより正確に検出することができることを確認した．

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７件）

① Hiroshi Ishii, Qiang Ma, Masatoshi Yoshikawa: ″Incremental Construction of Causal Network from News Articles,″ Journal of Information Processing, 20, pp.207-215, 2012. 査読有
② Xinpeng Zhang, Yasuhito Asano, Masatoshi Yoshikawa: ″A Generalized Flow Based Method for Analysis of Implicit Relationships on Wikipedia,″ IEEE Transactions on Knowledge and Data Engineering, 10 Nov. 2011. 査読有
③ 青戸 了, 清水 敏之, 吉川 正俊: ″整合性を考慮した注釈伝播,″ 日本データベース学会論文誌, Vol. 10, No. 1, pp. 49-54, 2011年６月. 査読有
④ Shin Ishida, Qiang Ma, Masatoshi Yoshikawa: ″Extraction of Characteristic Description for Analyzing News Agencies,″ Journal of Digital Information Management, Volume 8, Issue 6, pp. 349-354, December 2010. 査読有
⑤ 青戸 了, 清水 敏之, 増田 耕一, 吉川 正俊: ″履歴管理システムにおける多粒度アノテーション伝播,'' 日本データベース学会論文誌, Vol. 9, No. 1, pp. 64-69, 2010年６月. 査読有
⑥ Takeharu Eda, Masatoshi Yoshikawa, Toshio Uchiyama and Tadasu Uchiyama: ″The Effectiveness of Latent Semantic Analysis for Building Up a Bottom-up Taxonomy from Folksonomy Tags,″ World Wide Web Journal, Volume 12, Number 4, pp. 421-440, DOI http://dx.doi.org/10.1007/s11280-009-0069-1, December, 2009. 査読有
⑦ Umaporn Supasitthimethee, Toshiyuki Shimizu, Masatoshi Yoshikawa and Kriengkrai Porkaew: ″XSemantic: An Extension of LCA based XML Semantic Search,″ IEICE Transactions on Information and Systems, Vol. E92-D, No. 5, pp. 1079-1092, May 2009. 査読有

〔学会発表〕（計 32 件）

① Yu Suzuki and Masatoshi Yoshikawa, ″QualityRank: Assessing Quality of Wikipedia Articles by Mutually Evaluating Editors and Text,″ The 23rd ACM Conference on Hypertext and Social Media, Milwaukee, WI. USA. June 25-28, 2012.
② Damien Eklou, Yasuhito Asano, Masatoshi Yoshikawa, ″How the Web can help Wikipedia: A Study on Information Complementation of Wikipedia by the Web,″ The 6th International Conference on Ubiquitous Information Management and Communication (ICUIMC), Kuala Lumpur, Malaysia, February 20-22, 2012.
③ Jiyi Li, Qiang Ma, Yasuhito Asano, Masatoshi Yoshikawa, ″Ranking Content-Based Social Images Search Results with Social Tags,″ The Seventh Asia Information Retrieval Societies Conference (AIRS 2011), pp.147-156, Dubai, UAE, December 18-20, 2011.
④ Ling Xu, Qiang Ma, and Masatoshi Yoshikawa, ″Credibility-Oriented Ranking of Multimedia News Based on a Material-Opinion Model,″ The 12th International Conference on Web-Age Information Management (WAIM 2011), Wuhan, China, September 15, 2011.
⑤ Tetsutaro Motomura, Toshiyuki Shimizu, and Masatoshi Yoshikawa, ″Alternative Query Generation for XML Keyword Search and Its Optimization,″ 22nd International Conference on Database and Expert Systems Applications (DEXA 2011), Toulouse, France, August 30, 2011.
⑥ Kazuki Tawaramoto, Junpei Kawamoto, Yasuhito Asano, and Masatoshi Yoshikawa, ″A Bipartite Graph Model and Mutually Reinforcing Analysis for Review Sites,″ the 22nd International Conference on Database and Expert Systems Applications (DEXA 2011), Toulouse, France, August 31, 2011.
⑦ Ryo Aoto, Toshiyuki Shimizu, and Masatoshi Yoshikawa, ″Propagation of Multi-granularity Annotations,″ 22nd

International Conference on Database and Expert Systems Applications (DEXA 2011), Toulouse, France, September 2, 2011.
⑧ Xinpeng Zhang, Yasuhito Asano, Masatoshi Yoshikawa, ″Towards Improving Wikipedia as an Image-rich Encyclopaedia through Analyzing Appropriateness of Images for an Article,″ The 13th Asia-Pacific Web Conference (APWeb), Beijing, China, April 18-20, 2011.
⑨ Xinpeng Zhang, Yasuhito Asano, Masatoshi Yoshikawa: ″Enishi: Searching Knowledge about Relations by Complementarily Utilizing Wikipedia and the Web,″ The 11th International Conference on Web Information System Engineering (WISE), pp. 480-495, Hong Kong, China, December 12-14, 2010.
⑩ Ling Xu, Qiang Ma, Masatoshi Yoshikawa: ″Exploring Special Items in Multimedia News Based on a Stakeholder Model,″ IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT2010), pp. 452-455, Toronto, Canada, Aug.-Sep. 2010.
⑪ Umaporn Supasitthimethee, Toshiyuki Shimizu, Masatoshi Yoshikawa, and Kriengkrai Porkaew: ″Meaningful Interrelated Object Tree for XML Keyword Search,″ Proc. of the 2nd International Conference on Computer and Automation Engineering (ICCAE 2010), pp. 339-344, Singapore, February 27, 2010.
⑫ Yusuke Kiritani, Qiang Ma, and Masatoshi Yoshikawa: ″Classifying Web Pages by Using Knowledge Bases for Entity Retrieval,″ Proc. 20th Int. Conf. Database and Expert Systems Applications (DEXA2009), LNCS 5690, pp. 761-768, Linz, Austria, August 31 - September 4, 2009.
⑬ Takeharu Eda, Toshio Uchiyama, Tadasu Uchiyama and Masatoshi Yoshikawa: ″Signaling emotion in tagclouds,″ Proceedings of the 18th International Conference on World Wide Web (WWW 2009), poster, pp. 1199-1200, Madrid, Spain, April 20-24, 2009.
⑭ Xinpeng Zhang, Yasuhito Asano, and Masatoshi Yoshikawa, ″Visualized Elucidations of Ranking by Exploiting Object Relations,″ Third International Workshop on Ranking in Databases (DBRank 2009), Shanghai, China, March 29, 2009.
⑮ Mizuho Iwaihara, Kohei Murakami, Gail-Joon Ahn, and Masatoshi Yoshikawa, ″Risk Evaluation for Personal Identity Management Based on Privacy Attribute Ontology,″ 27th International Conference on Conceptual Modeling (ER2008), LNCS 5231, pp. 183-198, Barcelona, Spain, Oct. 2008.
⑯ Yumi Yonei, Mizuho Iwaihara, and Masatoshi Yoshikawa, ″Person Retrieval on XML Documents by Coreference Analysis Utilizing Structural Features,″ in Proc. 19h Int. Conf. Database and Expert Systems Applications (DEXA2008), Lecture Note in Computer Science 5181, pp. 552-565, Turin, Sep. 2008.
⑰ Qiang Ma and Masatoshi Yoshikawa, ″Ranking People Based on Metadata Analysis of Search Results″, LNCS 5176, Web Information Systems Engineering -WISE 2008 Workshops (E-BAG2008), pp. 48-60, Auckland, New Zealand, Sep. 2008.
⑱ Xinpeng Zhang and Masatoshi Yoshikawa, ″Querying RDF Data using Dynamic Concise Bounded Description,″ Intenational Workshop on Information-explosion and Next Generation Search (INGS 2008), Shenyang, China, April 26, 2008.
⑲ Umaporn Supasitthimethee, Toshiyuki Shimizu, Masatoshi Yoshikawa, and Kriengkrai Porkaew, ″An Extension of LCA based XML Keyword Search″, International Workshop on Information-explosion and Next Generation Search (INGS 2008), Shenyang, China, April 26, 2008.
⑳ Takashi Menjo and Masatoshi Yoshikawa, ″Trend Prediction in Social Bookmark Service Using Time Series of Bookmarks,″ WWW2008 Workshop on Social Web Search and Mining(SWSM2008), Beijing, China, April 22, 2008.

６．研究組織

(1)研究代表者

吉川　正俊（YOSHIKAWA MASATOSHI）
京都大学・大学院情報学研究科・教授
研究者番号：30182736

(2)研究分担者
馬　強（MA QIANG）
京都大学・大学院情報学研究科・准教授
研究者番号：30415856
浅野　泰仁（ASANO YASUHITO）
京都大学・大学院情報学研究科・特定准教授
研究者番号：20361157
清水　敏之（SHIMIZU TOSHIYUKI）
京都大学・大学院情報学研究科・助教
研究者番号：60402468
岩井原　瑞穂（IWAIHARA MIZUHO）
早稲田大学・理工学術院（情報生産システム研究科）・教授
研究者番号：40253538
鈴木 優（SUZUKI YU）
名古屋大学・情報基盤センター・研究員
研究者番号：40388111