科学研究費助成事業　　研究成果報告書

平成　２６　年　　６　月　１７　日現在

機関番号：　１４３０１
研究種目：　若手研究(B)
研究期間：　2010～2013
課題番号：　２２７０００９７

研究課題名（和文）半構造データに対する付加情報の管理と検索への利用

研究課題名（英文）Managing Additional Information for Semi-structured Data and its Application to Search

研究代表者
清水　敏之（SHIMIZU, Toshiyuki）

京都大学・情報学研究科・助教

研究者番号：６０４０２４６８

交付決定額（研究期間全体）：（直接経費）　　3,000,000円、（間接経費）　　900,000円

研究成果の概要（和文）：様々なデータおよび文書は半構造データとしてとらえることができ、その書式としてＸＭＬが広く利用されている。半構造データに関して、付加情報を取得、提示することによる理解支援に関する研究を行った。特に論文、プレゼンテーションスライド、科学データのメタデータを対象として、これらを半構造データとして扱い、細粒度での文書間対応付け、ＸＭＬキーワード検索の支援、多粒度注釈伝播などに関して手法を提案し、評価して有効性を確認した。

研究成果の概要（英文）：Many kinds of data or documents can be regarded and represented as semi-structured data, and XML is widely used for the format. We studied on understanding support for semi-structured data by retrieving and presenting useful information. In this study, we used academic papers, presentation slides, and metadata for science data, which can be regarded as semi-structured data. We proposed and evaluated some new methods for document alignment in fine granularity, supporting XML keyword search, propagation of multi-granularity annotations, and so on.

研究分野：　総合領域

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：　半構造データ　ＸＭＬ　理解支援　情報検索

１．研究開始当初の背景

ＸＭＬに代表される半構造データに関する研究は近年盛んに行われている。ウェブ上には大量の文書が存在するが、多くのものは半構造データとして扱うことが可能であると思われる。ＸＭＬは論理的には木構造であり、データを表形式で扱う従来の関係データベースに関する技術をそのまま利用することは一般的にはできない。ＸＭＬに関する研究としては、例えばノードに対するラベリング手法や索引を生成することによる検索の高速化手法などがある。

一方、近年の情報量の爆発的な増加を受け、データを理解するためのより洗練された仕組みが求められている。ウェブ上のデータに対するウェブサーチエンジンといった検索の仕組みを利用することで、膨大な情報の中から必要だと思われるものを取得することが可能ではあるが、検索対象データの増大につれて、検索結果の量も増加し、理解が困難になる。また、取得されたデータそのものに対する理解に関しては検索技術だけでは不十分である。

データに対する理解支援のためには、データそのものに加えて、データに関連する付加情報を利用することが有効である場面が多い。例えば、論文を読む際に、その内容を紹介するスライドを見ることは理解の補助になる(データ間の関連情報)。また、論文情報を管理するウェブサイトである DBLP では、論文書誌情報そのものに加え、著者別ページでは投稿先(会議や論文誌)ごとの投稿回数や共著者の回数などを提供している(統計情報)。ウェブ上の百科事典である Wikipedia では、現在の記事だけでなくそれまでの編集履歴を保持し、過去の版からの差分情報を提供することで、より高い信憑性を提供していると言える(履歴情報)。

２．研究の目的

データの多様化に際し、ＸＭＬのような半構造データに対する研究が近年盛んである。また爆発的なデータ量の増加から、データに関して深く理解することが困難になっている。本研究では半構造データを対象とし、データに対する付加情報を管理して検索に利用することで、データ管理者、データ利用者の双方に対しての理解支援を行う。

付加情報としては、データ間の関連情報、統計情報、履歴情報の３種に着目する。半構造データを扱う上で重要なデータ粒度に対する考察を行い、それぞれの付加情報を効率的に扱い、効果的に利用するための具体的なシステム構築と実験による評価まで行う。

３．研究の方法

(1) データ間の関連情報

あるデータに対して関連するデータが存在する場合、それらの対応付けを行う。半構造データを対象とすると、文書レベルだけでなくデータ内の要素レベルでの対応取得が考えられる。データとしては特に論文に着目する。節と段落の入れ子関係などの文書構造を意識して対応付けする手法を開発する。

(2) 統計情報

文書中の語の出現情報等を利用して、その文書を特徴付ける情報を取得し、より適切な検索結果の取得や、得られた検索結果の理解支援を行う。半構造文書では、部分文書ごとに異なる特性を持っている場合があり、細粒度での解析が求められる。

(3) 履歴情報

データの生成過程や処理過程に関する情報は、データに対する理解を深める上で有用な情報であり、データの品質、つまりデータの正確性や妥当性を示すために不可欠な情報である。出力データと起源データとの細粒度な対応を考慮し、起源データに付与された注釈を出力データに伝搬させる手法に関して考察する。

４．研究成果

(1) データ間の関連情報

① 論文とプレゼンテーションスライドの対応付け

論文およびプレゼンテーションスライドを半構造データとしてとらえ、会議で発表された論文と発表時に用いられたプレゼンテーションスライドを部分対応付けすることで理解補助につながると考え、図１のように細粒度で対応付けを行う手法を提案した。対応付け確信度の高い対応を軸として他の対応を補正し、さらに文書構造を利用して段階的に対応付けを行った。

実際に小規模なデータセットを用いて対応付け実験を行い、人手で判定した結果と比較して議論を行った。文書順および構造情報を利用することにより対応付け精度の改善を行うことができた。

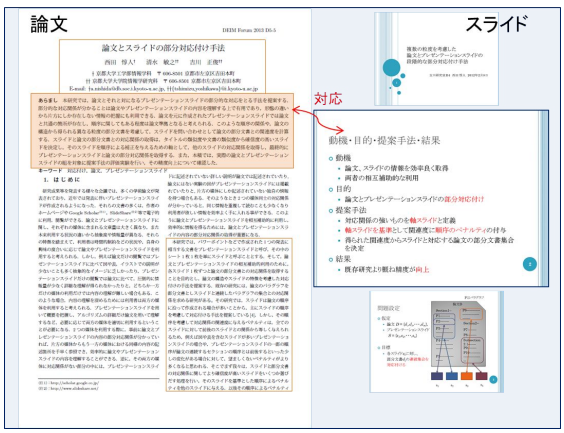

図１ 論文とスライドの対応付け

② 関連文書間の対応付け

同一の著者による国際会議論文と論文誌論文などのような同様のトピックに関する

複数の文書を細粒度で対応付ける手法を提案した。特に異言語の文書ペアでは内容による類似度だけでは対応付けが不明確な場合があり、文書順や文書の階層構造の利用を検討した。
段落単位の対応付けを想定し、文書構造を利用して段落の統合、入れ替えまで考慮して最終的な対応付けを得ることとした。複数の論文ペアを用いて実験を行い、構造を利用することによる対応付け精度向上を確認した。

③ 地球科学データと論文の関連付け
地球科学データを対象とし、データに関連する論文を提示することでデータ理解支援ができると考え、データと論文情報の関連付けを行う手法に関して考察した。関連付けを行うことにより、データと論文のそれぞれに対して情報が補強され、閲覧時の理解補助ができると考えられる。
実際のデータセットと論文の対応付け事例を調査して、メタデータを介して論文と対応付けるための手法を議論した。

(2) 統計情報
① キーワード検索結果の理解支援
文書集合に対するキーワード検索において、検索結果中から統計的な分析に基づいてキーワードを抽出し、検索結果を増分的に拡張する手法を提案した。拡張によって検索者は周辺情報を得ることができる。
また、文献データベースに対してキーワード検索を行う状況を考え、単に検索結果を提示するだけではなく、検索結果に関連する情報や、別の視点を得ることができる情報を提示する手法を考案した。具体的にはXMLデータに対してＬＣＡに基づくキーワード検索を行う際に、入力されたキーワード集合によって得られた検索結果と同一の結果を得ることができる別のキーワード集合(代替問合せ)を取得し、利用者に提示することによって理解支援を行う手法を提案した。
さらに、代替問合せの取得にあたり、その高速な処理アルゴリズムを提案し、実験による検証を行って優位性を確認した。このような代替問合せの提示は元の問合せと共通のキーワードが少ない際に特に有用であると思われるが、我々が提案する手法ではキーワード数が増加しても高速に処理可能であった。

② 要素の特性を考慮したXMLキーワード検索
ＸＭＬ文書に対するキーワード検索として、XML文書中の文書指向の部分、データ指向の部分を意識した検索結果の取得を行う手法に関して提案した。検索者の問合せ意図を推定し、要素間の繋がりを考慮して適切な検索結果の構築を行う。
ＸＭＬ文書におけるデータ指向要素および文書指向要素の混在に着目した点は新規性があり、XML化した論文のデータを用いて実験を行って有用性を議論した。

③ 地球科学データに対するキーワード推薦
地球科学データに対するメタデータに着目し、メタデータを解析することで、地球科学データに対して適切なキーワードを付与する手法を考案した。キーワード情報はデータの俯瞰・検索に重要である。
あらかじめ定まった語彙集合(統制語彙)からの付与を考え、単なる文字列マッチングだけでなく、機械学習を利用し、語彙集合内の語彙の階層構造を利用することで、より適切なキーワード付与を行うことができた。実際に地球環境情報統融合プログラム(DIAS-P)におけるメタデータを利用し、データ担当者による評価を行って有効性を確認した。

(3) 履歴情報
① 多粒度注釈伝播
データ品質や理解支援のため情報を保持する手段としてデータに対して細粒度のメタデータすなわち注釈を付与することが考えられる。注釈付与に際し、処理を通して付与された注釈を伝播させる手法および伝播された注釈の整合性分析に関して研究を行った。
関係データを想定し、多粒度で付与された注釈を伝播させるにあたり、処理の基となるそれぞれの関係演算に関して議論を行い、従来研究に比べた優位性を実験により確認した。
また、注釈を付与する際には根拠となるデータが存在する場合があると考え、関係演算の拡張を行うことにより、根拠データを注釈と同時に伝播させていくことで注釈の客観性を維持する注釈管理手法を提案した。

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２件）

① 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、査読有、Vol．10、No．1、2011、pp．49-54
http://dbsj.org/journal/dbsj_journal/dbsj_journal_vol_10_no_1_49_54/
② 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、査読有、Vol．9、No．1、2010、pp．64-69
http://dbsj.org/journal/dbsj_journal/dbsj_journal_vol_9_no_1_64_69/

〔学会発表〕（計２１件）

① Toshiyuki Shimizu, Tomo Sueki, and Masatoshi Yoshikawa, "Supporting Keyword Selection in Generating Earth Science Metadata," 37th Annual IEEE Computer Software and Applications Conference (COMPSAC 2013), Kyoto, Japan, July 25, 2013.
② 田邊 翼, 清水 敏之, 吉川 正俊, "キーワードの役割と要素の特性を考慮したXML検索," 第5回Webとデータベースに関するフォーラム(WebDB Forum 2012), 2012年11月20日, 東京.
③ Tsubasa Tanabe, Toshiyuki Shimizu, and Masatoshi Yoshikawa, "Effective Keyword-Based XML Retrieval Using the Data-Centric and Document-Centric Features," 8th Asia Information Retrieval Societies Conference (AIRS 2012), Tianjin, China, December 18, 2012.
④ 冨田 典也, 清水 敏之, 齊藤 昭則, 吉川 正俊, "重要度と時空間近接度を統合した地球科学データのランキング手法," 第4回Webとデータベースに関するフォーラム(WebDB Forum 2011), 2011年11月4日, 東京.
⑤ 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.
⑥ 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.
⑦ Masashi Tatedoko, Toshiyuki Shimizu, Akinori Saito, and Masatoshi Yoshikawa, "A Retrieval Method for Earth Science Data Based on Integrated Use of Wikipedia and Domain Ontology," 21st International Conference on Database and Expert Systems Applications (DEXA 2010), Bilbao, Spain, September 2, 2010.

６．研究組織

(1)研究代表者

清水　敏之（SHIMIZU, Toshiyuki）
京都大学・情報学研究科・助教
研究者番号：６０４０２４６８