

(続紙 1)

京都大学	博士（情報学）	氏名	河合 英紀
論文題目	A Study on Search Methods for Knowledge Extraction and Structuring from the Web (Webからの知識抽出・構造化のための検索手法に関する研究)		
(論文内容の要旨)			
本論文では、インターネット上の大量の情報の中からユーザが所望の知識を取得するために、情報収集のための2大タスクである情報検索技術（IR）と情報抽出技術（IE）を統合し、Webからの知識抽出・構造化を効率よく行うための検索方式について提案を行っている。 IRとIEの統合方法としては、大きく、Corpus-drivenとSearch-drivenの2つのアプローチに分類できる。Corpus-drivenアプローチでは、IEは大規模文書コーパスに直接アクセスして情報抽出を行い、その結果をIRのインデックスにメタデータとして付与する。この場合、情報抽出を行うためには大規模文書コーパスをローカルに構築する必要があるため、実現可能なベンダーが限定されてしまうという問題があった。一方、Search-drivenアプローチでは、IEはIRに対してクエリを発行し、得られた検索結果に対して情報抽出を行うため、大規模文書コーパスをローカルに構築する必要はなく、誰もが必要な情報抽出モジュールを設計できるという利点がある。 従来のSearch-drivenアプローチでは、ターゲットとなる知識が頻出する言語パターンをクエリとして用いるテンプレート型クエリが一般的である。しかし、従来研究では、ターゲットとなる知識の表現に多様な表記ゆれが存在する場合、クエリとして利用する言語パターンが膨大になり、検索エンジンに多大な負荷をかけてしまう問題があった。そこで本論文では、言語パターンの代わりにターゲットとする知識の例をクエリとして用いるインスタンス型クエリと、そのための情報抽出のアーキテクチャを提案する。インスタンス型クエリを用いた情報抽出では、多様な表記ゆれが存在するような知識抽出タスクにおいても網羅性の高い検索が可能である一方、テンプレート型クエリに比べて抽出された知識の候補にはノイズも含まれるため、本論文では、機械学習などを用いた判定処理の提案も行っている。また、本論文では、なるべく少ない検索回数で多くの知識を抽出するためのクエリ拡張方式もアーキテクチャの一部として提案した。 本論文は全6章から構成されている。その概要は以下の通りである。 第1章は序論であり、本論文の研究の背景、本論文の研究を行うに至った動機、および、本論文の研究の全体の概要について述べている。 第2章では、第3～5章において説明するそれぞれの研究課題に特に深く関連する従来研究について整理し、本研究の位置づけを明確にしている。			

第3章では、同一ドメインに属するオブジェクトの集合 (Homogeneous Object Set) を知識として抽出する例として、オブジェクト集合クエリによる同位語辞書構築方法を提案しこれについて論じている。提案手法はインスタンス型クエリとして、複数のシード語を用い、ヒットしたページ内におけるシード語周辺のローカルなパターンを使い捨てにしながら新語を抽出することを特徴としている。評価実験の結果、従来手法と比較して10倍程度の検索効率と繰り返し処理に対する高いロバスト性を有することを明らかにした。

第4章では、レコード構造のオブジェクト (Object Tuple) を知識として抽出する例として、ドメイン混合型クエリによる未来／過去イベント情報の検索方法を提案しこれについて論じている。提案手法では、イベントが起こる時間範囲を表す時間限定詞、特定の年度を表現する年度表現、未来／過去のイベントを表現する際に頻出するキーワードであるコンテキスト語、を組み合わせたクエリ拡張によって効率よくイベント情報を収集し、イベント情報中のキーワードと年度の意味的な対応関係を機械学習で判定する。評価実験の結果、過去のイベント情報抽出には時間限定詞のような言語パターンを用いたテンプレート型のクエリ拡張が有効であるが、未来のイベント情報抽出には、年度表現やコンテキスト語のようなインスタンス型のクエリ拡張が有効であることが明らかになった。

第5章では、オブジェクト間の関係性を表すネットワーク (Heterogeneous Object Network) を知識として抽出する例として、動向表現クエリによる因果関係ネットワークの抽出方法を提案しこれについて論じている。提案手法では、政治、経済、社会、技術等に関する動向の要因となるキーファクターを、変化や傾向を表す動向表現を用いたクエリ拡張によって再帰的に収集し、キーファクター同士の間の因果関係の有無を機械学習で判定する。評価実験の結果、因果関係を表す言語パターンをクエリ拡張に使う従来手法に比べて、網羅的で高精度な因果関係ネットワークを抽出できることを示した。

第6章では、本研究で得られた研究成果をまとめ、さらに今後の展開について述べている。

(論文審査の結果の要旨)

本論文は、大規模な文書コーパスを持たないユーザであっても、検索エンジンを活用することによって、同一ドメインのオブジェクト集合、オブジェクトのレコード構造、オブジェクトのネットワーク構造など、様々な形態の知識の抽出・構築を可能にする検索方式を提案し、その有効性の検証を行ったものである。

従来、情報検索を用いた情報抽出方法としては、言語パターンをクエリとして用いるテンプレート型クエリが用いられていた。しかし、従来手法では、ターゲットとなる知識の表現に多様な表記ゆれが存在する場合、クエリとして利用する言語パターンが膨大になり、検索エンジンに多大な負荷をかけてしまう問題があった。検索回数と得られる知識の網羅性はトレードオフの関係にあるため、より少ない検索回数でより網羅的な情報を抽出することは困難であるが非常に重要な課題である。

学位申請者は、言語パターンの代わりにターゲットする知識の例をクエリとして用いるインスタンス型クエリと、そのための情報抽出のアーキテクチャを提案している。提案アーキテクチャにおいては、言語パターンよりも緩い条件で網羅的に知識の候補を収集するクエリ拡張方式と、機械学習を用いて知識とノイズを分離する判定方式が重要な構成要素である。さらに、提案手法の有効性検証のための評価実験を行い、興味深い結果を提示している。

具体的には、本論文の研究によって得られた成果は以下のように要約される。

1. 同一ドメインに属するオブジェクトの集合を知識として抽出する例として、オブジェクト集合クエリによる同位語辞書構築方法を提案した。提案手法は複数のシード語の集合をクエリとして検索を行い、ヒットした文書群からシード語周辺の一貫性の高いパターンを有する構造化文書を選定し、構造化文書中のパターンを使い捨てにしながら新語を抽出することを特徴としている。従来手法と比較して約10倍の検索効率と、繰り返し処理に対する高いロバスト性を有することを実験的に示した。
2. オブジェクトのレコード構造を知識として抽出する例として、ドメイン混合型クエリによる未来／過去イベント情報の検索方法を提案した。提案手法では、時間限定詞、年度表現、コンテキスト語を組み合わせたクエリ拡張によって効率よくイベント情報を収集し、イベント情報中のキーワードと年度の意味的な対応関係を機械学習で判定する。評価実験の結果、過去のイベント情報抽出には時間限定詞のような言語パターンを用いたクエリ拡張が有効であるが、未来のイベント情報抽出には、年度表現やコンテキスト語のようなインスタンス型のクエリ拡張が有効であることを示した。
3. 多様なオブジェクト間の関係性を表すオブジェクトのネットワーク構造を知識として抽出する例として、動向表現クエリによる因果関係ネットワークの抽出方法を提案した。提案手法では、政治、経済、社会、技術等に関する動向の要因となるキーフアクターを、変化や傾向を表す動向表現を用いたクエリ拡張によって再帰的に収集し、キーフアクター同士の間の因果関係の有無を機械学習で判定する。評価実験の結果、従来手法に比べ網羅的で高精度な因果関係知識を抽出できることを示した。

以上、本論文は、情報検索と情報抽出を統合するための新たなフレームワークを提案し、その中で検索効率の高いクエリ拡張方式と、ノイズを除去するための判定方式を提案しており、学術上、實際上寄与するところが少なくない。よって、本論文は、博士（情報学）の学位論文として価値あるものと認める。また、平成23年8月11日に論文内容とそれに関連した事項について諮問を行った結果、合格と認めた。