IPA

2008 年度下期未踏 IT 人材発掘・育成事業　採択案件評価書

1. 担当PM

竹田　正幸 PM（九州大学大学院　システム情報科学研究院　教授）

2. 採択者氏名

チーフクリエータ：　渡部　浩昭（インペリアルカレッジロンドン　計算機学科　研究員）
コクリエータ：　なし

3. プロジェクト管理組織

株式会社ゼータ

4. 委託金支払額

5,200,000 円

5. テーマ名

機械学習システムを用いた Web 空間からの知識発見

6. 関連Webサイト

なし

7. テーマ概要

「予測」は、科学、経済学、社会学など幅広い適用分野を持つ基盤技術である。
精度の高い予測を行うためには「エラーの少ないデータ収集」「専門家によるデータ解析」「予測結果の妥当性の確認」が必要になり、その結果多大なコストが通常発生

することから、「気象予測」「株価予測」などの分野で限定的に利用されている現状がある。

このような「テーラーメード」の予測とは別に、予測精度は多少落ちてでも、大量に存在する Web 空間上のテキストデータから(半)自動的にモデルを構築し積極的に予測技術をサービスとして利用していきたいという要望が存在する。

本提案は、このような要望に応えるために「自然言語処理技術と機械学習アルゴリズムを用いた Web 空間からの知識発見ソフトウェア」の開発を行い、汎用予測システム実現に向けたプロトタイプシステムの構築を提案する。

本開発の技術的な独自性は、

（1）自然言語処理技術を用いた「定性的な知識」の抽出、

（2）頻度情報を用いた「定量的な知識」の抽出、

（3）確率的帰納論理プログラミングを用いた記号的統計学習

の 3 点にある。

確率的帰納論理プログラミング（Probabilistic Inductive Logic Programming: PILP）は、確率論理(Probabilistic Logic)を記述言語として不確実性を含む知識の帰納的学習を実現する。PILPはサポートベクターマシンと組み合わせて予測精度を向上させたり、生物学分野において「ロボット科学者」の頭脳部分として用いるなど、理論・応用の両面で発展を続けているが、近年「関係」を「三つ組み」として表記するセマンティック Web との親和性の高さから、Web 空間からの知識発見問題への適用も期待されている。なお、帰納論理プログラムを用いて予測を行う場合、背景知識の「質」と「量」が予測精度に多大な影響を与える。背景知識の構築はドメインエキスパートの協力のもとで慎重に行なわれるが、時間とコストの観点からボトルネックとなってきた。

上記のような背景の下で、今回の開発では Web 上のテキスト情報から自然言語解析技術を用いて知識を自動抽出することにより、背景知識の構築に関わるボトルネックを解消する。構築した知識の「質」を向上させるためにドメインエキスパートのアドバイスを仰ぐ際には、自動構築した背景知識を「たたき台」として用いることによりコスト削減に貢献できる可能性が高い。

なお、ユーザーはキーワードを用いて「予測」を行う問題領域を変更できるため、基盤技術に必要な柔軟性も同時に備えたプロトタイプシステムの開発を目指す。

８. 採択理由

帰納論理プログラミング(ILP)に基づく機械学習システムを用いてWebからの知識発見に役立てようとする提案である。理論計算機科学分野の成果である ILP 技術を現実の問題に適用しようとする際のボトルネックに正面から取り組もうとするもので、高い未踏性を有する。開発計画も具体的かつ明瞭であることから、この開発計画は着実に進行するものと判断し、採択とした。

９. 開発目標

Web 上で情報蓄積が進むに伴い、様々な分野において網羅的な知識が入手可能になっている。例えば、経済新聞の記事データベースには「専門家による解説」として実際に生起した「事象」およびその考えられる「原因」が、文章形式で大量に蓄積されている。我々の調査によると、ある経済新聞データベースには過去５年の記事の中で「因果関係」を含むものが 14 万件以上（１日平均 70 件以上）存在した。ここで、発生した事象を「結果」、専門家が挙げた原因を「原因」とすると、経済記事データベースは経済分野における「因果関係データベース」と見なせる。この因果関係データベースは「予測」の観点から非常に興味深い。なぜなら、「原因 A」と「結果 B」からなる「A → B」という因果関係を知っている場合、新たに「A」が発生した際に、「B がその結果として生起する」と「予測」できるからである。さらには、「B → C」という因果関係を知っている場合には、「A」が発生した際に、因果関係の連鎖を用いて「C」の生起の「予測」が行える。

Web 上の情報をデータベースとして捉え推論を適用する試みは、「セマンティックWeb」と呼ばれる研究分野ですでに提唱されている。セマンティックWebの根本的な難点は、不完全知識を完全知識にするための知識の補完作業が人間に委ねられている点である。つまり、情報の欠損により不明な点があった場合、その欠損を人間が埋めない限り不明なままである。このセマンティック Web の持つ「不完全知識の下での推論」問題は、過去に人工知能分野の「エキスパートシステム」研究が直面した問題と本質的に同じである。

機械学習は、この不完全知識の下での推論問題に対して、既知の情報から未知の知識を「仮説」として生成することにより不完全な知識を補完する方法論を提供した。本プロジェクトもこの「機械学習的アプローチ」を Web 上の情報を用いた予測システム構築に適用する。

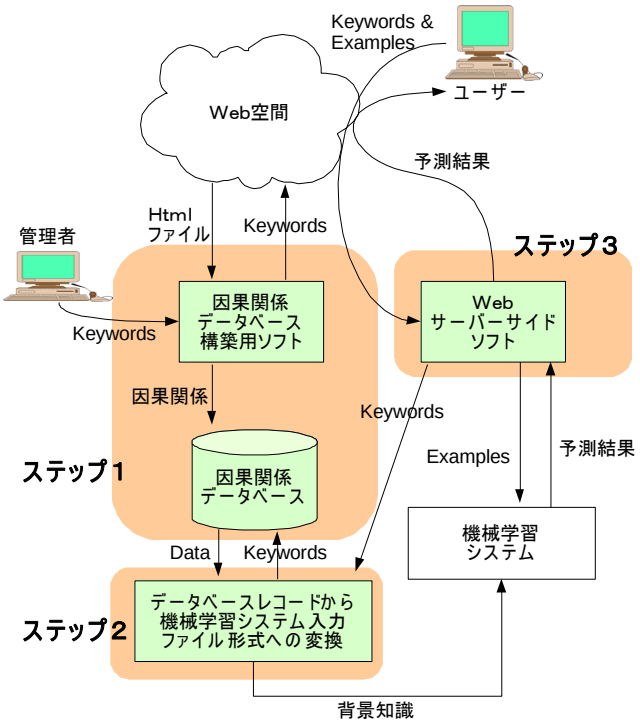
本プロジェクトでは、Web 空間のテキストファイルから自動抽出した因果関係から、「仮説的」因果関係を生成するツールを開発し、不完全な知識の下でもユーザが予測を行えるようにすることを目的とする。

10. 進捗概要

本個別プロジェクトでは、機械学習の一手法である帰納論理プログラミングシステムを利用することにより、不完全な因果関係知識に対して補完機能を持つ予測システムの構築を行った。開発した予測システムは、

- 因果関係知識の収集および整備を行う「予測サービス提供者」へのツール群
- 不完全な知識を補完する「ユーザ」へのツール群

からなる Java プログラムである。



システム全体図

上図は今回作成したシステムの全体図を示している。ステップ1、2、3に含まれるソフトウェアモジュールはJava言語を用いて開発し、機械学習システム部は既存の帰納論理学習システムを利用した。今回開発したシステムは、インターネットに接続されたコンピュータ2台から構成される。

ソフトウェアの開発は、以下の 3 ステップにて行った。

- ステップ1：因果関係データベース
  インターネット上のニュースサイトから自動抽出した因果関係を保持するデータベース構築
- ステップ2：知識表現変換ソフトウェア
  ユーザーからの要求に応じて因果関係データベースレコードを帰納論理学習システムへの入力ファイル形式に変換するプログラム開発
- ステップ3：Web サーバサイドソフトウェア
  既存の帰納論理学習システムをインターネットサービスとして利用するためのインターフェースソフトウェア開発

11. 成果

本個別プロジェクトで開発した成果物は、以下の4つに大別される。

（A）ユーザインタフェース
（B）因果関係データベースシステム
（C）知識表現変換ソフトウェア
（D）Web サーバサイドソフトウェア

このうち、（B）(C)については、特許出願及び論文投稿を予定しているため、委託業務期間完了後5年間は秘匿ノウハウに指定し、非公開とする。

以下では、（A）（D）についてその概要を述べる。

（A）ユーザインタフェース

≪一般ユーザ向けインタフェース≫

一般ユーザは、Web ブラウザを通して予測システムへとアクセスする。各タブを選択することにより、以下の7つの機能が利用できる。

1. ページ検索機能タブ： データベースに保持されている Web ページを検索する機能。ユーザは「キーワード」およびデータの格納されているディレクトリを指定することで検索を行う。
2. 因果関係検索機能タブ： キーワードを入力すると、そのキーワードを含む Web ページを検索し、該当 Web ページ内に含まれる因果関係「のみ」をユーザに表示する機能。
3. 入力ファイル生成タブ： 機械学習システムへの入力ファイルを作成する機能。ユーザがキーワードとデータの格納ディレクトリを入力すると、「因果関係検索機能タブ」で検索されたのと同様な因果関係が内部的に抽出され、その後機械学習システムへの入力ファイル形式に変換される機能。
4. 予測機能タブ： ユーザが保存している「仮説的因果関係」を含むファイル名と、事実集合を列挙した「モデル」ファイル名を指定すると、予測を行う機能。

5. 仮説生成タブ： 因果関係文（英語文）を入力すると、「仮説因果関係」を生成する機能。
6. キーワード検索タブ：「原因部」「結果部」各々に含まれるキーワードを指定すると、該当する因果関係を検索し表示する機能。
7. 連鎖情報生成タブ： まず原因部 A に含まれるべきキーワードを指定すると、該当する因果関係 A→B がまず表示される。次に B に含まれるキーワードをユーザがリンクをクリックする形で指定すると、B→C なる次の因果関係が表示される。これを繰り返すことにより因果関係の連鎖を検索できる。

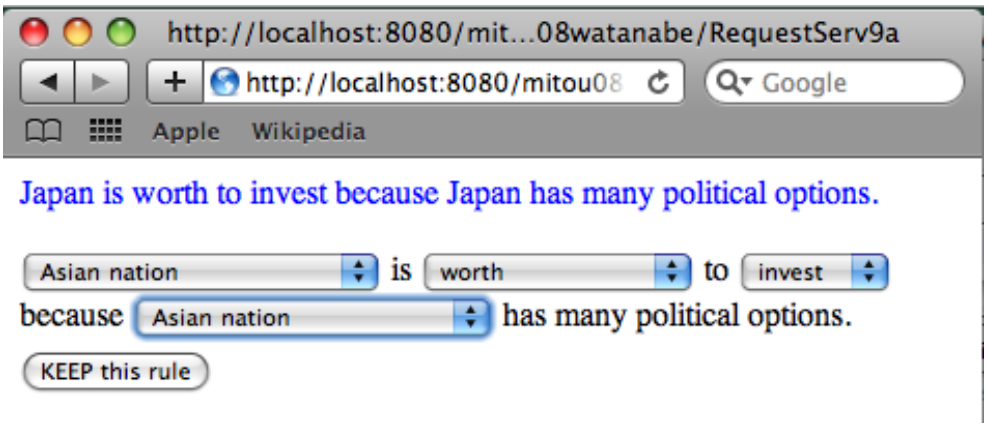
≪管理者用インタフェース≫

1. ファイル変換タブ： Web ページが格納されているディレクトリを指定すると、そのディレクトリに含まれる HTML ファイルから HTML タグを取り除き、拡張子 WB を持つページのテキストファイルを作成する機能。
2. インデックス作成タブ： 全文検索用のインデックスを作成する。データが格納されたディレクトリを指定する必要がある。
3. インデックス検索機能：「インデックス作成タブ」にて生成されたインデックスを用いて全文検索を行う機能。

他のタブはユーザ向け機能と同様である。

≪候補仮説表示インタフェース≫

システムからユーザへと提示される候補仮説の数は莫大になる可能性があるため、コンパクトにユーザに候補仮説を表示する必要がある。本開発では、プルダウンメニューを効果的に利用することにより、コンパクトに仮説をユーザに提供するインタフェースを開発した。例えば、「日本は政策オプションが存在するため、投資に値する」という因果関係を「仮説生成タブ」上のツールから入力すると、システムは下図のように仮説として「アジアの国々には政策オプションが存在するので投資に値する」という仮説的因果関係を画面表示する。プルダウンメニューには他の多くの選択肢が残っているが、ユーザはその中から 1 つを選択し、「Keep this rule」ボタンを押すことでその仮設をファイルに保存することが可能なインタフェースになっている。



(D)Web サーバサイドソフトウェア

以下の2つの機能を実装した。

- Web サーバ機能： ユーザへのサービスをインターネット経由で提供するためのサーバサイド機能。ユーザからキーワード等のパラメータ値を受け取り、予測結果を HTML ファイル形式で返す Java サーブレットプログラムを作成した。
- 内部インタフェース機能： ユーザから得た情報を各内部モジュールへと受け渡すインタフェース機能、および機械学習システムから予測結果を受け取る機能を実装した。

12. プロジェクト評価

Web 上の情報をデータベースとして整備し推論を適用する試みは、セマンティックWebと呼ばれる研究分野ですでに提唱されている。しかし、セマンティックWebの枠組みでは記事内に書かれている情報を用いての予測を行うことになるので、知識の拡大は発生せず、下図において「Web 情報空間」と示される「書かれている知識」の範囲内に留まっている。

本プロジェクトにより実現されたツールを用いると、上記で「Web 仮説空間」と示されている、通常の広く利用されている既存のサーチエンジンなどではアクセスできない、仮説的因果関係を用いた推論により獲得された「仮説知識空間」が構築できる。この「Web 仮説空間」は、既存の「検索」ベースの Web サービス企業が提供する情報空間とは全く異なるという「新規性」を持つ。

13. 今後の課題

本個別プロジェクトにおいて予定していた開発項目はすべて実現した。今後の課題としては、以下の2つがあげられる。

（1）日本語を含む他言語への対応。

（2）予測精度情報のユーザへの提供。

このうち、（1）について、現在は英語のみの対応となっているが、日本語を含む他言語への対応が望まれる。使用している自然言語処理ライブラリはドイツ語と中国語の取り扱いが可能であるため、この 2 言語への対応は比較的容易に実現できるが、日本語への対応はこれからの研究の進展を待たねばならない。

一方、（2）については、現在の仕様では、候補仮説はコンパクトにユーザに提示された後、ユーザが自ら仮説を選択する仕様になっている。今回の開発物は、定性的な予測を行うために、定量的な情報は利用していないが、何らかの定量的な情報が

利用可能であれば、候補仮説からの仮説の自動選択やユーザへの予測精度情報の提供が可能になる。

本開発成果物は、検索エンジンに基づいたインターネット空間の探索というアプローチとは異なり、自らインターネット空間を拡大させながら探索を行うという新しいタイプのインターネットの利用方法を提案している。有料の予測サービス以外にも、生成する仮説空間を共有することにより他者とのコミュニケーションの発生が生じ、コミュニティの生成が期待できるなど、いろいろな側面を考えながら今後の展開を進めていく必要がある。