機械学習システムを用いた Web 空間からの知識発見
―自然言語処理と帰納論理学習を用いたテキスト情報からの仮説生成―

1. 背景

Google などのサーチエンジンは、「存在する」莫大な量の情報から必要な情報を探し出す。Wolfram Alpha などのナレッジエンジンは、「存在する」事実に関する質問に回答する。では手つかずの「存在していない」情報をユーザーに提供するアプリケーションはあるのか？

予測の分野には「ある」。 予測分野では存在する事実から存在していない情報を「仮説」として生成する方法が広く議論されている。中でも生成したモデルに基づいて予測を行う方法は科学、経済、社会的問題領域で広く用いられている。

精度の高い予測を行うためには，「精度の高いデータ」を「ある一定量」集めたうえで「専門家による解析と妥当性の確認」が必要になることから多大なコストが現状では発生するため、予測技術は個人レベルで気軽に利用できるレベルには至っていない。だが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手軽に予測技術を利用していきたいという要望も確実に存在している。

我々の調査によると、ある英文新聞紙にはモデル生成に有用な情報を含む専門家による経済記事が過去5年に14万件程度（一日平均70件以上）存在した。このような一般的に入手可能な情報を網羅的、動的に集める事によって従来の予測問題では扱えなかったような予測を行うサービスを比較的安価に実現出来る可能性が増している。

2. 目的

サーチエンジンでは扱えない仮説的 Web 空間を生成しユーザーに提供するために

- 「テキストファイルから予測に有用な情報を自動抽出しモデル生成を行うツール
- 作成したモデルを用いて予測支援をおこなうためのツール

を含む予測システム HypoWeb の開発を行う事を本プロジェクトの目的とした。

図1：HypoWeb ユーザーインターフェース

3. 開発の内容

上記目的を達成するために（1）自然言語処理モジュール、（2）仮説生成用帰納論理プログラミングモジュール、（3）推論（予測）モジュールからなるシステムを Java 言語にて開発した。全てのサービスは Web サービスとしてユーザーとサービスクリエータに提供される。

一般ユーザー向けサービス（図1）

一般ユーザーは，Webブラウザを通して予測システムへとアクセスし「情報収集」「モデル作成」「予測」を行う。ユーザーは合計7つのツールが利用できる。

1. ページ検索機能：データベースに保持されているWebページをキーワード検索する機能。
2. 予測用知識検索機能：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予測に有用な知識「のみ」をユーザーに表示する機能。
3. 入力ファイル生成機能：機械学習システムへの入力ファイルを作成する機能。
4. 予測機能：ユーザーが作成した仮説と，事実集合を列挙したモデルから予測を行う機能。
5. 仮説生成機能：予測用知識を自然言語文で入力すると，「仮説」を生成する機能。
6. キーワード検索機能：予測用知識を詳細にキーワード検索する機能。
7. 連鎖情報生成機能：連鎖的な予測を行うために必要な予測用知識を検索する機能。

サービスクリエータ向けサービスは、上記機能に加えてデータベース構築用ツールを提供する。

仮説表示インターフェース：マニュアルモード(図2)

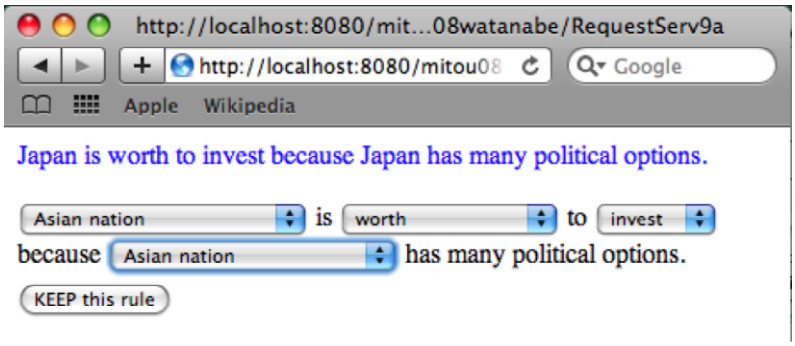

図2：仮説表示インターフェース（マニュアルモード）

システムからユーザー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ーに「仮説」候補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ーに提供するインター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」という予測用知識を「仮説生成機能」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，ユーザーはその中から 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ーフェースになっている。

なお、上記のようなユーザーとのインタラクションは行わずに、ユーザーの入力したキーワードに基づいて自動的に仮説を生成する自動仮説生成モードも提供する。

4．従来の技術（または機能）との相違

データマイニング、情報抽出の分野ではテキストからの情報獲得を行うソフトウェアが数多く提供されているが、開発者の知る限り、Web 上のテキストから自然言語処理を行って自動獲得した「人が読む事の出来る」文章を利用して仮説を自動生成し、予測を行う機能をもったソフトウェアおよびサービスは存在していない。

既存のサーチエンジン（Google, Yahoo, Bing など）は、埋もれている知識を探し出して来てユーザーに提示するが、仮説を生成しているわけではない。Wolfram Alpha は自然言語処理と推論を組み合わせたサービスを提供するが、これも既知の知識を用いているため仮説を生成する我々のプログラムとは明らかに異なっている。

5．期待される効果

開発したソフトウェアは、高精度の予測を実現する「オントロジー」「Web コンテンツ」に価値があることを示している。Web 空間のコンテンツの新たな評価基準を提供しているため、本ソフトウェア開発を通して「予測向けコンテンツ作成」産業分野の創成を目指している。

6．普及（または活用）の見通し

現在二通りのサービス提供を模索している。

1. 企業内に存在する書類と社内データベースを用いたビジネスユーザー向けの予測サービス
2. 開発したソフトウェアを用いた Web 上での一般ユーザー向け予測サービス

特に後者の予測サービス提供に向け現在準備を進めている。

7．開発者名（所属）

渡部浩昭（インペリアルカレッジロンドン　計算機学科）

（参考）開発者URL

http://www.doc.ic.ac.uk/~hw3/