# 研 究 報 告 書

## 「ネットワーク理論と機械学習を用いたウェブ情報の構造化・知識化」

研究期間: 平成 20 年 10 月～平成 24 年 3 月
研 究 者: 松尾 豊

1，研究のねらい

本研究では，ネットワーク理論と機械学習に基づく，画期的なウェブ情報の統合・知識化アルゴリズムの構築を目指す。特に，エンティティ(人物や組織，物質，製品名等)のネットワークに着目し，目的に応じた予測のための構造化技術を構築する。大量のウェブ情報に書かれたエンティティ間の構造を抽出し，目的に応じて知識として利用するための基盤であり，ウェブの次世代「知識エンジン」につながる技術である。

2，研究成果

研究成果としては、1)予測に基づく知能の実現の仕組みに関する思考、2) 具体的なアルゴリズムとしての実装、3) ウェブ情報をマイニングすることによる社会予測の研究、の大きく3つに分けられる。以下ではそれらを順に説明する。

研究の動機は、「考える脳、考えるコンピュータ」で語られているような予測に基づく知能のアルゴリズムを構築したいということである。現状の機械学習における各手法では、訓練データとして十分に工夫した素性に基づくデータを与えるとうまく学習でき、そうでなければうまく学習できない。機械学習のパフォーマンスを決める最も重要な部分は、人間がよって行われる適切な素性の構築であり、アルゴリズムでうまく行うことができない。一方、人間は何らかの方法で素性の構築を行っており、このことと人間が世界を構造化して捉えていること（エンティティとその関係性という認識をしていること）、ならびにその構造化が次に何が起こるかをより的確に予測するために用いられているということが相互に密接に関連しているのではないか。こうした研究の動機から出発し、領域会議等でのさまざまな議論や文献等の調査、それに基づく思考を経て、この問題に対する重要な要素として以下のものを抽出した。

1. ニューラル・ダーウィニズム：予測の精度が高いニューロン群だけが生き残る。
2. 補助問題の生成：訓練データそれ自身の予測を行うことで、便宜的に問題数を増やし、訓練データの数を増やす。その後、素性の転移等によって、より効率的な学習を行う。
3. ネットワークと中心性：素性間のネットワークを構築した際に、他の素性と関連の強い（中心性の高い）素性は、さまざまな問題に有用なロバストな素性である。また、中心性が高いと、多くの他の素性にアンカーされることになり、動きにくい
4. 言語化：こうしたネットワークにおける部分集合（サブグラフ）が、それを指すポインタとしての何らかの言語表現と紐付けられ、任意のタイミングで活性化することが可能となる。

次に、これをアルゴリズムとして構築する方法についての成果を述べる。ここで構築するアルゴリズムは、自然言語の文書を入力することで、予測をベースにして自動的にシラブルや単語がチャンク化されるアルゴリズムである。入力したデータの一部（テキストの場合、文字や単語など）に由来する「センサーノード」と、その発火を予測する「予測ノードの」の2種類を使うことで、

入力したデータのチャンク化、抽象化等を行うものである。言語によるラベルは、この2部グラフのクラスタとの関連で作られることになる。

以下のような一連の手続きでネットワークが構築され、それが予測に使われる。

1. ひとつの文字に対応するセンサーノードを設定する。各センサーノードに対する予測ノードを作る。
2. 共起が高い2つのセンサーノードに対して、センサーノードと予測ノード間にエッジを張る。
3. センサーノードの発火を学習させる。予測ノードへのほかのセンサーノードからのエッジの重みを修正する。
4. エッジの重みが重いものだけ残す。
5. 予測ノードと、予測ノードに対応したノードが同時に on になったら(予測に成功したら)発火するセンサーノードを作る。2 に戻る。

1 は準備フェーズであり、2 のフェーズは、あり得る素性の解空間を効率的に探索するためのヒューリスティックである。3 は、通常のニューラルネットワークや SVM 等での重み(判別関数)の学習である。4 は、枝刈りのフェーズである。5 が一番特徴的な点であり、予測に成功したことを表すノードを新たに作ることで、重要な文脈は保存されていくことになる。そしてこの文脈自体がセンサーノードとなって、新たにそれを予測するような一連のノード群が構成されていく。

これを具体的に可視化したものが図1である。大規模な文書データを入力することで、次の文字を予測するために、シラブル、単語に該当するノードが現れ、相互に連結されている様子が分かる。

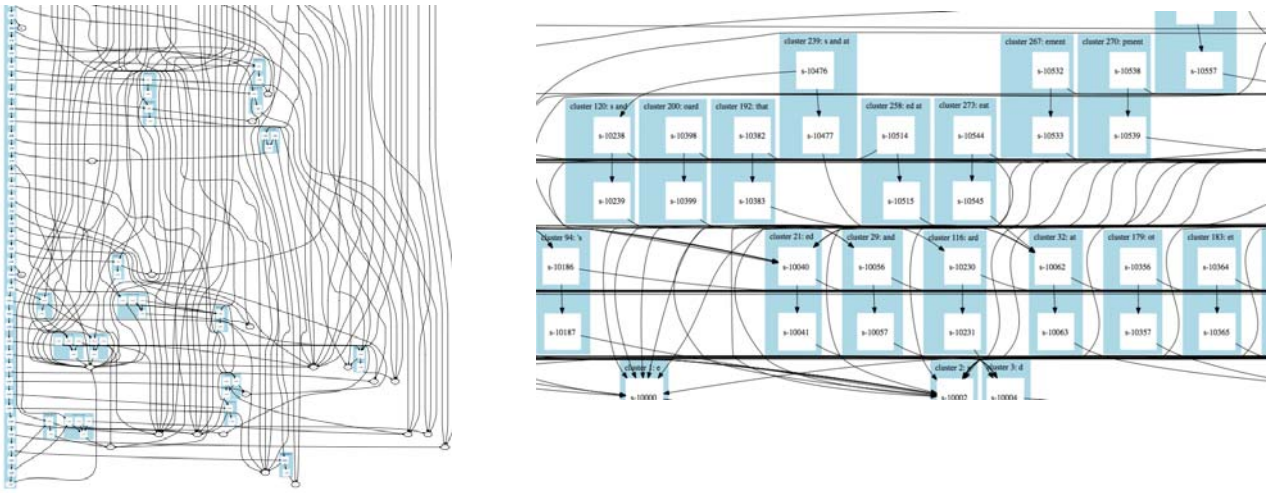

図 1 文書から得られたシラブル、単語のネットワーク構造

これが下図のように階層をもったネットワークとして構造化されていくことが期待できる。また、特定の2部グラフの構造を他のノードが指すことにより、概念を指すポインタとして機能し、それによってより高次の抽象化や関係概念の発見が可能になる。

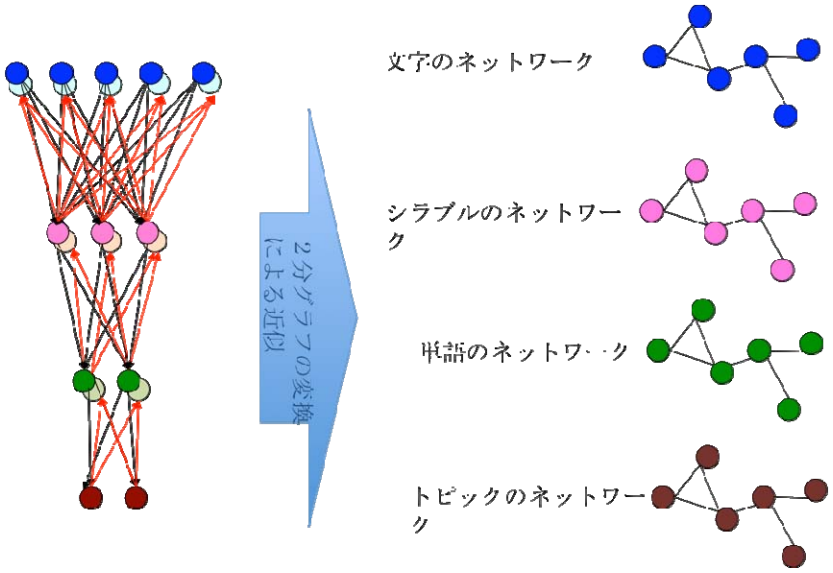


図 2 階層間の2部グラフと各階層におけるネットワーク

本研究では、提案するアルゴリズムをウェブ上のデータに適用するための試みとして、ウェブ上の情報に対して機械学習を用いて予測を行う応用例をいくつか示した。ひとつは、ブログ上の情報からの選挙予測、もうひとつは、twitter からの地震の早期検出の研究である。

下記は国内のブログデータから、国政選挙に関するブログを抜き出し、過去の国政選挙の結果とそのときのブログの特徴（候補者名および党名に対する言及の程度と内容）から選挙の得票数を予測しようという試みである。この結果、ブログ上での言及と得票数は高い相関にあり、精度の高い予測が可能であることが明らかになった。

| | # predictions | # correct | # wrong | # unsure | Hit ratio |
|---|---|---|---|---|---|
| **Blog analysis (ours)** | 254 | 241 | 39 | 36 | 92.33% |
| Asahi Shimbun | 260 | 245 | 15 | 40 | 94.62% |
| Nikkei Shimbun | 288 | 264 | 24 | 12 | 91.67% |

図 3 ブログからの選挙結果の予測の精度

また、下記は、twitter におけるツイートから、地震に関するツイートだけを抽出したものである。ツイート中に含まれるキーワードやその出現位置等を素性として用いることで、地震に関するツイートを一定の精度で抽出することができ、早期に検知することができる。この成果は WWW2009 において論文が採択され、その後、同様の研究における先進的な取り組みとして国内外にインパクトを与えることができた。


図 4 地震に関するツイートの抽出

以上、本研究では、予測に基づく知能の実現の仕組みに関する思考と、それに基づくアルゴリズムの構築、さらにウェブ情報からの社会予測の応用例を示した。相互の課題が直接的な基礎−応用の関係になっているわけではないが、本研究で目的としていたネットワーク構造を用いた画期的な学習、およびその先に広がる世界の像を少なからず示すことが出来たのではないかと考えている。

## 3, 今後の展開

本研究の今後の展開としては2つの方向がある。ひとつは、提案したアルゴリズムの評価である。アルゴリズムは、データから構造を抽出しながら予測精度の向上を目指すもので、これまでの機械学習のアルゴリズムと比べ、パラメータが多い。この部分の最適化とともにアルゴリズムの評価をしていく必要がある。

もうひとつは、さまざまな社会予測（選挙やイベント）という観点での応用であり、その際には、予測したい対象とは（一見）関係がないものの予測に寄与する素性を抽出することで、本来予測したい対象の予測に寄与するという枠組みが有用であることが想定される。本アルゴリズムを活用する形でこういった社会予測への応用を進めて行きたいと考えている。

## 4, 自己評価

提案する手法に関して議論を深め手法の大枠を示すことができた点、またウェブマイニングに関してインパクトのある応用例を実現できた点では当初の目標を達成したと考える。一方で、アルゴリズムとしての精錬や評価の点ではまだ課題が多く、当初目標通りではない。設定したテーマが難しい課題であり、3年間でどこまでの成果が上がるかは想定しにくいところではあったが、アドバイザーの先生方等の意見は非常に参考になり、研究助成を受けなかったら進められなかったであろう部分まで大きく進めることができた。大きなチャレンジの土台にあたる部

分であり、今後この期間の研究が花開くように努力していきたい。

5，研究総括の見解

ネットワーク理論と機械学習を用いたウェブ情報の構造化・知識化についての研究である。ウェブマイニングにおける「エンティティ」発見をネットワーク構造上の機械学習の観点から捉えていて独創的であり、ウェブマイニングの核となるアルゴリズムを発見・構築したいという意気込みに期待していた。

提案する手法に関して議論を深め、手法の大枠を示すことができ、また、ウェブマイニングに関して社会的にもインパクトのある応用例を実現し、公表している。これらは彼ならではの優れた成果と考える。また、当初の目標を達成しており、この点も高く評価したい。一方で、構想自体は大きなものであるため新しい今後の研究課題も多く出ている。アルゴリズムとしての精錬や評価の点では、今後、更なる成果につながっていくことを期待する。

6，主な研究成果リスト

（1）論文（原著論文）発表


| |
|---|
| 1. Yutaka Matsuo and Hikaru Yamamoto: Community Gravity: Measuring Bidirectional Effects by Trust and Rating on Online Social Networks, Proc. 18th International World Wide Web Conferenc (WWW2009), 2009 |
| 2. Danushka Bollegala, Yutaka Matsuo and Mitsuru Ishizuka: Relational Duality: Unsupervised Extraction of Semantic Relations between Entities on the Web, Proc. 18th International World Wide Web Conference (WWW2010), 2010 |
| 3. Takeshi Sakaki, Makoto Okazaki, and Yutaka Matsuo: Earthquake Shakes Twitter Users: Real-time Event Detection by Social Sensors, Proc. 18th International World Wide Web Conference (WWW2010), 2010 |
| 4. 岡 瑞起，松尾 豊: 検索エンジンを用いた関係の重みづけ，人工知能学会論文誌，Vol. 25, No. 1, 2010 |
| 5. Yingzi Jin, Ching-Yung Lin, Yutaka Matsuo, Mitsuru Ishizuka: Mining Longitudinal Network for Predicting Company Value. IJCAI 2011: 2268-2273 |