1B2-6

ネットワーク作る君：Webマイニングを用いたネットワーク自動抽出システム

Mr. Network Extraction: A Network Extraction System from the Web

石田 啓介[*1] 松尾 豊[*1*2] 安田 雪[*3]
Ksuke Ishida Yutaka Matsuo Yuki Yasuda

[*1]産業技術総合研究所 National Institute of Advanced Industrial Science and Technology
[*2]スタンフォード大学 Stanford University
[*3]東京大学 University of Tokyo

Mr. Network Extraction is a Web mining system to extract a network of a given set of entities. A user can input an arbitrary set of entities, e.g., researcher names, firm names, and celebrity names, and the system outputs a network among the entities. If two entites appear in many Web documents, the two entities are considered to be closely related thus become tied by an edge. In this paper, we overview the motivation, the concept and the system overview of the Mr. Network Extraction.

1. はじめに

Web 上の情報を活用する研究が進んでいる。ユーザが欲しい情報は、Web ページ自体であるとは限らず、したがって、検索技術を利用しながら情報を統合してユーザの質問にピンポイントで答える質問応答などの研究や、Web からトレンドを把握しようとする研究が行われている。

なかでも、エンティティとその関係性を抽出する研究は、ここ数年、徐々に活発に行われるようになった [4]。Web 上で言及されているエンティティとその関係は、Web ページ間の関係というレイヤーよりも上位のレイヤーの意味的な情報であり、そのための Web マイニング技術は、これまでの AI 技術の活用を考える上で、また今後の Web 上の情報処理を考える上で重要な位置を占める。

一方で、Web における SNS やブログといったソーシャルなシステムは、新たに社会ネットワーク分析の分野の興味も惹きつけている [**?**]。ますます多くの情報が Web 上で取得可能になり、それを利用して従来不可能であった大規模なネットワーク分析や新たな社会現象の分析を行うことも可能になりつつある。人工知能学会を対象にした研究者ネットワークの分析もここ 4 年ほど継続して行われている [3]。これは、Web や AI と社会学（特に社会ネットワーク分析）の融合分野の研究と考えることができる。

Web マイニングに関する研究の活用を考える上で、Web マイニングで得られた結果をさまざまな方法で簡単に利用できるようにすることは重要である。例えば、社会学の分野で Web 上の情報に着目した分析を行いたいとき、社会学の分野の研究者が必ずしもアルゴリズムを実装することが可能なわけではない。Web 上の情報の重要性、その統合の重要性を鑑みると、誰もが簡単に Web 上の情報を統合する（特にネットワークを抽出する）ことができるシステムを提供することは、研究の社会的なインパクトを高めることにつながる。

本稿では、Web マイニングに関して行われてきた研究 [1] を、誰もが手軽に利用できるようにするシステム「ネットワーク作る君」の概要を紹介する。

連絡先: 石田 啓介，産業技術総合研究所，〒 135-0064 東京都江東区青海 2-41-6　臨海副都心センター 426，03-3599-8294，03-3599-8255，ksuke@be.to

2. 抽出のアルゴリズム

「ネットワーク作る君」では、エンティティのリストを受け取って、その間のネットワークを抽出する。まず、エンティティごとに、検索エンジンのヒット件数を求める。（エンティティaに対するヒット件数を$hit(a)$とする。）さらに、エンティティのペアに対するヒット件数を求める。（$hit(a,b)$とする。）このとき、エンティティ間の関係の強さは、

$$hit(a,b)/(hit(a)+hit(b)-hit(a,b))$$

で与えられる。これは Jaccard 係数の場合であるが、他にもさまざまな係数で計量が可能である。

研究者ネットワークを抽出する場合には、関係のタイプを抽出することも可能である。これはあらかじめ学習した分類器を用いて、2 人の名前でヒットしたページを分類する。ネットワーク作る君では、共著、同研究室、同プロジェクト、同発表という 4 種類に分類する。詳しくは、[**?**] を参照されたい。

3. システムの概要と動作例

図 1 はネットワーク作る君のシステム概要を示している。ネットワーク作る君は、インターフェイス部は PHP、抽出部は Perl で記述された Web アプリケーションである。ユーザが入力する項目は、1. ネットワークの元となるノードのリスト、2. データ抽出に関する若干の設定、3. ネットワーク図の見栄えを調整する閾値などの設定である。ここでは簡単に、利用の流れを追いながら、システムの概要を説明する。

3.1 ノードデータの入力

ユーザはまず、ネットワークを抽出したいノードのリストを入力する必要がある。ここで入力するデータ形式は、ノード名とそのノードに関するキーワードのリストである。例えば、研究者のネットワークを抽出したい場合、ノード名は氏名、キーワードはその研究者の所属名を用いることが多い。実際の入力例を挙げると以下のようになる。

入力例

安田雪, 東京大学
松尾豊, スタンフォード大学
石田啓介, 産業技術総合研究所
....., ...

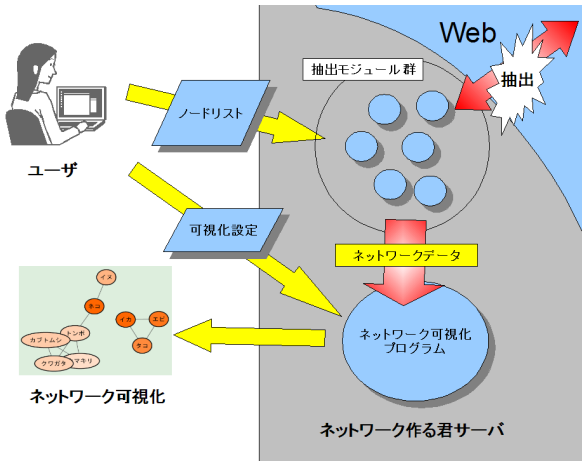

図 1: システム概要

図 2: ネットワーク抽出設定画面

3.2 ネットワーク抽出設定

ネットワーク抽出設定画面では、抽出処理に関する設定を行う。ヒット件数を求める際に利用する検索エンジンの選択、ネットワーク中心性の算出・関係タイプの抽出を行うか否か、などが現在選択可能である (図 2)。今後、ユーザによる関係タイプの定義、同姓同名問題を回避する手法の導入、研究者の所属名推定など、さまざまな抽出手法をここで選択できるようにする予定である。

設定の入力を終えると、サーバ上では抽出処理が開始され、設定に応じたモジュールプログラムが順番にコールされる。その間、ブラウザ上では Ajax により処理状況が逐次表示される。処理にかかる時間は、入力したノード数に比例して長くなるため、大規模なネットワークを抽出するような場合は、抽出完了メール機能により抽出完了のお知らせを受け取ることができる。受信したメールの本文中に URL が記述されており、そのURL にアクセスすることで続けて次のネットワークの調整を行うことができる。

3.3 ネットワーク図の調整

ネットワークデータの抽出が終了すると、画面はネットワーク図の調整へと移る。まずはじめに表示される画面では、何も調整を行っていないため、見やすいネットワーク図が得られているとは限らない (図 3)。ユーザに理解しやすいネットワーク図にするにはいくつかの調整が必要である。まず、共起ヒット件数・Simpson 係数・Jaccard 係数で閾値を設定し、弱い関係を非表示にすることで、強い関係が際立って表示されるようになる。また、関係の強さを Simpson 係数、Jaccard 係数など、さまざまな尺度に切り替えることができ、ネットワーク中心性の高さ、単独ヒット件数の多さをノードにグラデーションをつけることで表現できる (図 4)。これらの設定を変更するたびに、画面推移を行わず逐次ネットワーク図が変化するため、微調整をスムーズに行うことができる。ネットワークの生成には Graphviz を使用し、出力形式は SVG・Flash から選択可能である。

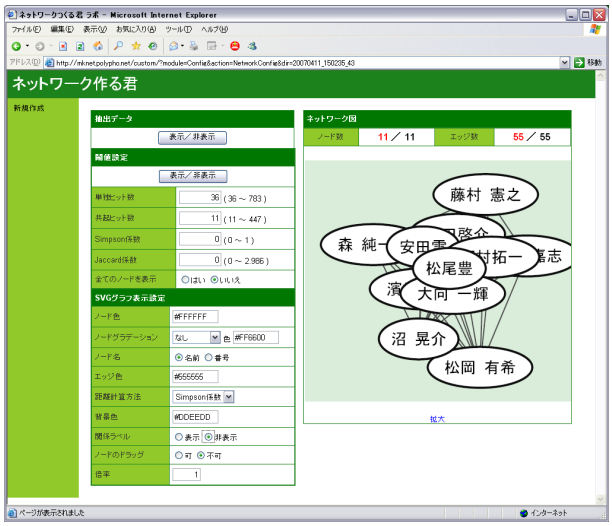

図 3: ネットワーク表示設定前

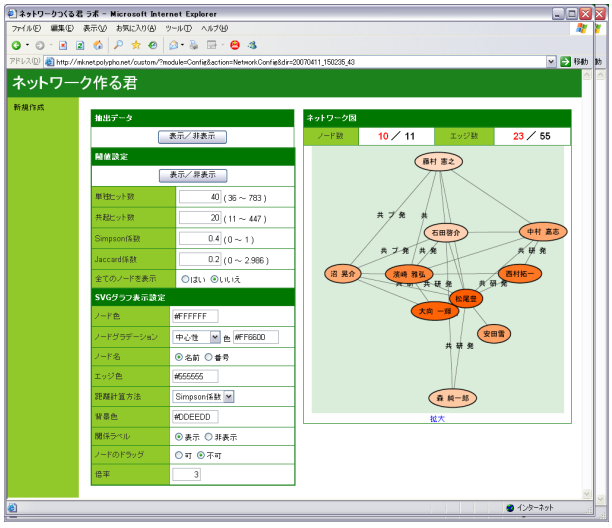

図 4: ネットワーク表示設定後

3.4 利用例

ネットワーク作る君で抽出できるデータは、もちろん研究者ネットワークだけではない。図 5 は、生物名 (カタカナ) によるネットワークである。見て分かるように、動物・植物・昆虫・鳥類というようにそれぞれが近寄って配置されている。図 6 は、戦国武将 20 人によるネットワークである。こうして得られたネットワークデータを、CSV 形式、DL 形式などさまざまな形式で出力することができる。

4. まとめ

本稿では、「ネットワーク作る君」を構築した背景と目的、その動作例、システムアーキテクチャー等について紹介した。今後、ユーザのニーズに応じて、多様な機能を提供していきたいと考えている。その結果、社会学を初めとしてさまざまな分野での分析、他のシステムでの活用につながっていければ、著者らの幸いとするところである。ネットワーク作る君は、http://mknet.polypho.net/から利用可能である。

参考文献

[1] Y. Matsuo, J. Mori, M. Hamasaki, H. Takeda, T. Nishimura, K. Hasida, and M. Ishizuka. POLYPHONET: An advanced social network extraction system. In *Proc. WWW 2006*, 2006.

[2] M. Pasca, D. Lin, J. Bigham, A. Lifchits, and A. Jain. Organizing and searching theworldwideweb of facts - step one: the one-million fact extraction challenge. In *Proc. AAAI2006*, 2006.

[3] 安田 雪, 松尾 豊, and 武田 英明. 人工知能学会におけるネットワーク構造と変化. In **人工知能学会全国大会**, 2006.

[4] 松尾 豊. 世界へのインタフェースとしての検索エンジン. **電子情報通信学会誌**, 2007.

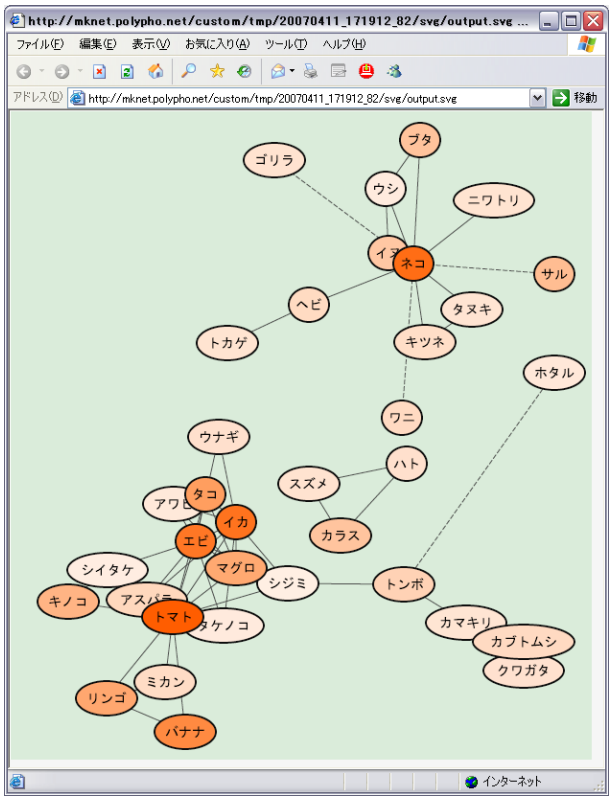

図 5: 生物名 (カタカナ) によるネットワーク

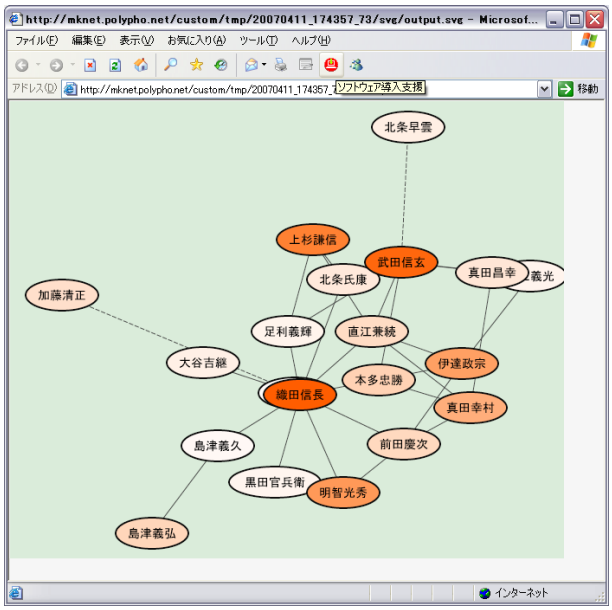

図 6: 戦国武将 20 人によるネットワーク