1F1-02

Webからの人間関係ネットワークの抽出と情報支援

Social Network Extraction from the Web

松尾 豊*1 友部 博教*2 橋田 浩一*1 石塚 満*2
Yutaka Matsuo Hironori Tomobe Kôiti Hasida Mitsuru Ishizuka

*1産業技術総合研究所 サイバーアシスト研究センター
National Institute of Advanced Industrial Science and Technology

*2東京大学大学院情報理工学系研究科
Graduate School of Information Science and Technology, The University of Tokyo

In a ubiquitous computing environment, it is desirable to provide a user with information depending on a user's situation, such as time, location, user behavior, and social context. At conventions, such as academic conferences and exhibitions, where participants must register in advance, the social context of participants can be extracted from th Web using their names and affiliations without asking them many questions. In our study, we attempt to extract the social network of participants from the Web, where a node represents a participants and an edge represents thbe relationship of two participants. The network shows the position of each participant and some participant clusters. Furthermore, this network can be used in many services, such as finding appropriate person to introduce or negotiate with someone, and who one should talk to in order to expand his/her network efficiently.

1. はじめに

学会や講演会などのイベント空間 [11] では、「人」が主役である．論文や資料を電子的に入手し後で目を通すのではなく、多くの人が多忙な時間を割いて実際に学会会場に足を運ぶ．これは、単に研究内容を理解するだけではなく、研究者が実際に話すのを聞いて、研究の背景や状況をよりよく把握するためである．さらに，会場の盛り上がりを肌で感じたり、多くの研究トピックを概観することができる．そして、他の研究者と会って話をしたり飲みに行くことによって、新しい情報を入手し、意見を交換し、次の研究や実用化の芽が生まれていく．

このように「人」が主役の場では、やはり「人」の関係が重要である．例えば、初対面の人と会った場合には、自分とその人とのつながりや、興味、共通の知人などが分かれば、コミュニケーションがよりスムーズに進むだろう．また、学会というコミュニティ全体の人間関係を見れば、どういう主要なグループがあり、自分はどこにいて、どういうグループの人とはあまり知り合いがいないのか理解できれば役に立つだろう．自分とある程度近い人の発表を推薦してあげればユーザはうれしいかもしれないし、人間関係は遠いけれども近い内容の研究をしている人との出会い支援は効果的かもしれない．

このように、ユーザの社会的な人間関係は非常に重要なユーザの文脈のひとつであると考えられるが、これまで情報支援の研究において着目されることは少なかった．それは、ユーザの人間関係の情報をシステムが獲得することが困難であるためである．例えば、各ユーザが自分と知り合い関係の人を明示的に入力し、知り合い関係のネットワークを構築する枠組み [2] も提案されているが、ユーザにとっては大変な作業である．

一方、Web マイニングの研究分野では、従来から、Web のリンク関係から重要なページを発見したり [1]、あるトピックに関する Web 上のコミュニティ [4][5] を発見する研究が行われている．近年では、特定の２人の人間をつなぐ知り合い関係のパスを抽出したり [5]、参照の共起性からコミュニティを発見する [6]、またあるページの評判情報を抽出する研究 [9] も行われている．

本研究では、ユーザ間の人間関係ネットワーク*1を Web から自動的に抽出する手法を提案する．検索エンジンを用いて人間関係のつながりの強さとその種類を判断する．なお、本システムは、本年度人工知能学会全国大会の学会支援システムにおいてサービス提供を行う予定である．

2. 人間関係ネットワークの作成

実世界やネットワーク上には多くの組織やコミュニティが存在し、それぞれに人間関係のネットワークがあると考えられる．本稿では、特に人工知能学会を対象にその人間関係ネットワークを抽出するが、他の学会やさまざまなコミュニティにも応用が可能である．

2.1 ノードとエッジの作成

さて、人間関係ネットワークは、ノードが人、エッジが人間関係を表すネットワークである．まず、ノードについては、全国大会のプログラムから入手した参加者（発表者）をノードとして設定した．さらに、できるだけ網羅的な関係を知りたいので、過去４年の全国大会に参加した人も、ノードとして設定した.（これも Web で入手できる情報である.）各ノードには、氏名に相当するラベルがつけられる．なお、氏名の他に所属情報も大会プログラムから抽出しているが、本システムで用いているのは参加者の氏名と所属情報だけである．

次に、ノード間にエッジを張る．ここでは、検索エンジン*2による検索ヒット数に基づいて２つのノードの関係の強さを測り，それに応じてエッジを張る．２人の氏名（X と Y）を検索クエリーとしたときの検索ヒット数が，偶然よりも多ければ，その２人の関係は強いと判断できる．つまり、X と Y に人間関係があるなら、"X and Y"で検索されたページには、それぞれのホームページや業績のページ、研究室のメンバーリスト

連絡先: 松尾 豊, 産業技術総合研究所 サイバーアシスト研究センター, 〒 135-0064 東京都江東区青海 2-41-6, Tel. 03-3599-8214, e-mail y.matsuo@carc.aist.go.jp

*1 社会ネットワーク（ソーシャルネットワーク）と呼んでもよいが、社会心理学におけるこの用語はより多様な関係を含んでいるため、ここでは謙虚に、人間関係の一部を表すネットワークという意味で、人間関係ネットワークと呼ぶことにする．

*2 Google.

表 1: 属性と値

属性	説明	値
NumCo	ページ内での X と Y の（同文内）共起回数	zero, one, or more_than_one
SameLine	X と Y が同じ行に出現しているかどうか	yes, or no
FreqX	X の出現回数	zero, one, or more_than_one
FreqY	Y の出現回数	zero, one, or more_than_one
GroTitle	語群（A〜F）がタイトルに含まれるかどうか	yes or no (for each group)
GroFFive	語群（A〜F）が最初の５行にあるかどうか	yes or no (for each group)

のページ、委員会や研究会などのページなどが含まれ、ヒット件数が多くなる．このような Web ページにおける氏名の共起関係の強さは、例えば共起頻度や相互情報量，Jaccard 係数などで測ることができる[*3]．

基本的に，我々が用いたのは Jaccard 係数を改良したものである．"X and Y"をクエリーとしたときの検索ヒット件数を $\#(X \cap Y)$、"X or Y"をクエリーとしたときの検索ヒット件数を $\#(X \cup Y)$ とすると，Jaccard 係数は，

$$rel(x,y) = Jaccard(X,Y) = \frac{\#X \cap Y}{\#X \cup Y} \qquad (1)$$

となる．ここで、$rel(x,y)$ はノード x と y の関係の強さを表す．全てのノードの組に対してこの値を計算し[*4]、$rel(x,y)$ が与えられた閾値を越えたらエッジを張る．

基本的にはこのように簡単な手法であるが、精度を上げるために次のような改良を行う．

- Web 上では同姓同名の人が多くいるため、目的とする人以外の関係が抽出されてしまったり，本来の関係が弱められてしまう可能性がある．そこで、名前に加えて所属情報も用いることで検索精度を上げる．所属機関に関して、複数の所属がある場合や，所属の変更がある場合、また所属機関に複数の名称や略称がある場合があるため，次のような工夫を行った．例えば、氏名が N、所属が A と B と C である場合には、" N and (A or B or C) " を検索クエリー X として用いる．たとえば、" 松尾豊 " の場合には、検索語は" 松尾豊 and (産業技術総合研究所 or 産総研 or 科学技術振興事業団 or 東京大学) "となる[*5]．

- 一般的に、Jaccard 係数は有名な人物のノードからは、あまりエッジを生成しない．これは、有名な人物ほどサーチエンジンによる検索件数が多くなるため、分母となる $\#X \cup Y$ が分子となる $\#X \cap Y$ に比べ非常に大きくなるためである．そこで、$\#X \cup Y$ を $min(\#X, \#Y)$ とする．ただし、このままでは、逆に検索ヒット件数が少ないマイナーな人物ほど分子が小さくなり，値が高くなりやすいので、最終的に次のような式を用いた．

$$rel(x,y) = \begin{cases} \frac{\#(X \cap Y)}{\min(\#X, \#Y)} & if\ \min(\#X, \#Y) > k, \\ 0 & otherwise. \end{cases} \qquad (2)$$

k は定数である[*6]

表 2: 語群

語群	単語
A	業績 論文 成果 発表 活動 テーマ 著者 publication papers themes activities awards 等
B	グループ 研究室 研究所 チーム メンバー 講座 member laboratory lab team institute etc
C	プロジェクト 委員会 project committee
D	研究会 定例 講演 報告 共催 会議 セミナー シンポジウム ワークショップ 開催
E	学会 プログラムじゃーなる 大会 全国 セッション 目次 開会
F	教授 専攻 教員 院生 講師 大学院 研究生

$rel(x,y)$ が閾値を越えた場合、ノード x と y の間にエッジを生成する．このエッジの長さは、関係が強くなるほど二つのノードの距離が短くなるように、関係の重みの逆数とする．ここで用いている式は，基本的には Jaccard 係数を改良した式であるが，相互情報量で得られた方向つきエッジの距離の短い方によりエッジのあるなしを決めているとも解釈することができる．

人間関係ネットワークは，(i) 図示してユーザに提示する，(ii) 情報支援に用いる，という 2 つの用い方があるが，(ii) の場合には，閾値を定めてノード間にエッジを張るかを必ずしも決める必要はない．(i) の場合には，ユーザに見やすいようにエッジの数を調整することが必要である．例えば，本システムでは，検索件数が多い有名な人物同士のエッジは現実的には重要であると考えられるので、Jaccard 係数に関わらず、$\#X \cup Y$ が閾値以上（70 件）を越えるエッジは表示するようにしている．

3. 機械学習による人間関係の分類

さて，人間関係ネットワークにおいて 2 つのノードを結ぶエッジが短いほど二人の関係が強いわけだが、実際の人間関係には関係の強さだけでなく、「同僚」や「委員会のメンバー」といった種類がある．各エッジにその関係の種類を表すラベルを付加することで、人間関係ネットワークの利用可能性が広がると考えられる．例えば「同僚」である 2 人より、「同じ研究会の発表者」である 2 人の方が関係が弱いだろうし，論文を推薦する際に同僚の論文を推薦するのは少しナンセンスである．このような人間関係の種類は、検索のヒット数だけでは捉えられない．そこで、本節では、人間関係の種類の判別するために、検索されたページの内容を見て、機械学習により得られたルールで判別する手法について述べる．

まず、それぞれのエッジに付加するラベル（クラス）を次のように定める．重複可能である．

- 共著：共著関係である．
- 研究室：同じ研究室や研究所のメンバーである．

*3 Jaccard 係数は、Referral Web[3] や [6] などでも用いられている．

*4 検索エンジンの利用規定を守るのが大変である．何日もかけてやらないと怒られてしまう．

*5 略称に関しては「東京大学＝東大」のように事前にシステムに登録しておく．

*6 本システムでは 30 とした．

表 4: クロスバリデーションによる評価

ラベル	共著	研究室	プロジェクト	発表
エラー率 (%)	4.1	25.7	5.8	11.2

表 5: 適合率と再現率

ラベル	共著	研究室	プロジェクト	発表
適合率 (%)	93.9	56.3	85.7	84.7
再現率 (%)	91.2	60.0	46.2	62.1

- プロジェクト：同じプロジェクトや委員会のメンバーである．
- 発表：同じ研究会や全国大会で発表している．

人間関係ネットワークを作成する際に，“X and Y” をクエリーとして検索ヒット数を得るが，その検索上位 3 ページを取得する．それぞれのページから表 1 のような属性の値を抽出する．語群は文書を特徴づけるものとして選択したものである[*7]．その語群を表 2 に示す．これらの属性とクラス（人間関係のラベル）を与え、C4.5[8] の帰納学習によって判別ルールを獲得する[*8]．

本研究では、人間関係ネットワーク生成のために収集したページのうち、ランダムに選んだ 275 ページに対して，正解となる人間関係のラベルを人手で与えた．得られた判別ルールの例を表 3 に示す．例えば，最も簡単なルールは，もし二つの名前が同じ行で出現していれば、共著関係と判断するというものである．獲得した判別ルールを用いてエッジにラベルを付加する．

5 分割のクロスバリデーションによる評価を表 4 に示す.「研究室」ラベルに関してエラー率が高いものの、全体的には判別精度が良い．また、表 5 にラベル判別の再現率と適合率を示す[*9]．

4. 例

図 1 に，本年度および過去 4 年間の人工知能学会全国大会の参加者間の人間関係ネットワークを示す．実際には 1500 人程度のノード数を持つグラフとなるが，全体の図示は困難であるので，中心的な（ヒット件数の多い）約 150 人からなるネットワークを示している．中心的なクラスタと，周辺のノードがある様子が分かる．

図 2 に，一部を拡大したものを示す．それぞれのノードは参加者に対応しており，エッジには「共著」「研究室」「プロジェクト」「発表」のラベルがつけられている．距離が近いほど，関係が強いと判断されていることを示している．

社会ネットワークには，ある領域全体を見るソシオセントリックネットワークと，ある個人を中心に見るエゴセントリックネットワークという分類がある [10]．図 1，2 は，ソシオセントリックネットワークと考えられるが，図 3 は 1 人の参加者を中心にしたエゴセントリックネットワークである．特定の参加者を中心に、そこから 2 エッジで到達できる参加者を表示している．

*7 語群は、判別ルールの訓練データを用い，クラスごとの頻出語から選択している．

*8 他の学習アルゴリズムでもよいが，解釈の容易性からまずは C4.5 を用いている．

*9 評価に用いたのは、人手で判別した約 500 ラベルであり，判別ルール作成には使用していない．

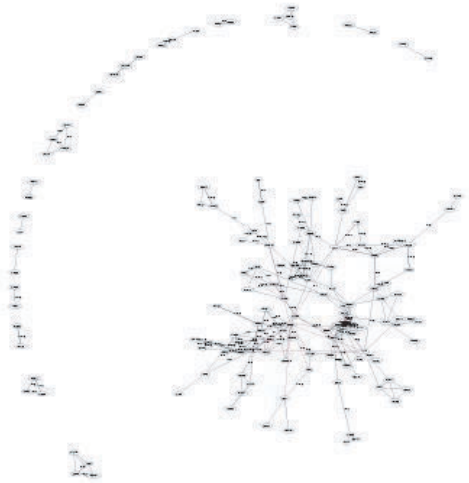

図 1: 人間関係ネットワーク全体

5. 議論

本研究は，Web マイニングという観点から [3][6] などの研究と近い．しかし，所与の名前と所属情報のリストから，その間の関係を網羅的に取り出すという点，ページの内容まで踏み込んで関係を判断している点で新規性が高い．

もちろん，Web ページだけからでは推測できない関係も存在する．例えば、Web 上に情報がないほど新しい関係や，フォーマルに Web に現れないような関係（例えば友人関係），昔の共著や所属の関係は Web から取得するのは難しい．しかし，現在でも人間関係を近似的に取り出すには十分な情報が Web 上にあり，また，今後ますます多くの情報が電子的に利用可能になると思われる．

ただし，いくつか考慮しなければならない点がある．検索エンジンの利用に関する問題[*10]，検索した文書の著作権の問題，ユーザのプライバシに関する問題など，社会的な合意がまだ不十分で難しい問題ではあるが，サービス提供の際には十分な配慮を行うことが必要である．

6. おわりに

近年、多くのコンピュータやセンサーが環境や機器に埋め込まれ、多様な情報通信インフラがシームレスに接続されるユビキタスネットワークに関する研究が行われている．特に、ユーザの位置情報、活動情報、欲求、嗜好、これまでの履歴などを考慮し、ユーザの文脈に応じた情報支援を行うことが、大きな課題の 1 つである [12][7]．基本的に、ユーザが入力した情報やセンサー情報によりユーザの文脈を推定することが試みられているが、ユーザを取り巻く人間関係も非常に重要な文脈のひとつであろう．

*10 細田 IT 担当相は，2003 年 4 月 18 日の衆院個人情報保護特別委員会で，検索エンジンについて「特定の個人情報を検索することができるように体系的に構成したものではないので，本法案の『個人情報データベース等』には該当しない」と述べ，検索エンジンの利用者や運営者は，個人情報保護法の規制対象とならないとの認識を示したが，現在のところこれが定着した見解とは言えない．

表 3: 判別ルール

ラベル	ルール
共著	SameLine = yes
メンバー	NumCo = more_than_two & GroTitle(D) = zero & GroFFive(A) = not_zero & GroFFive(E) = not_zero GroFFive(B) = not_zero & GroFFive(C) = zero FreqX = more_than_two & FreqY = more_than_two & GroFFive(A) = not_zero & GroFFive(D) = zero
プロジェクト	SameLine = no & GroTitle(B) = zero & GroFFive(F) = not_zero GroTitle(C) = not_zero
発表	FreqY = one & GroFFive(D) = not_zero & GroFFive(F) = zero GroTitle(A) = zero & GroFFive(B) = zero & GroFFive(D) = not_zero

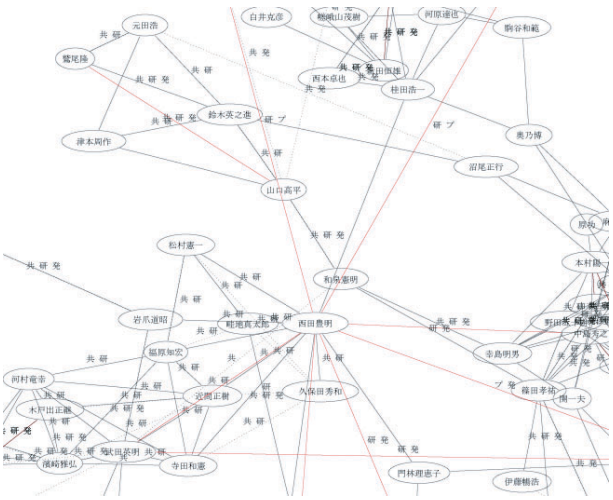

図 2: 人間関係ネットワークの一部

本稿では，人間関係ネットワークを Web から抽出する手法を提案したが，例えば，どういう人間関係の人が近くにいるときにはどういう支援を行えばよいのか，人間関係ネットワークの位置とユーザの活動状況の関係はあるのか，活発な学会はどのような人間関係ネットワークはどういう特徴をもっているのかなど，今後，さまざまな研究が可能であると考えられる．

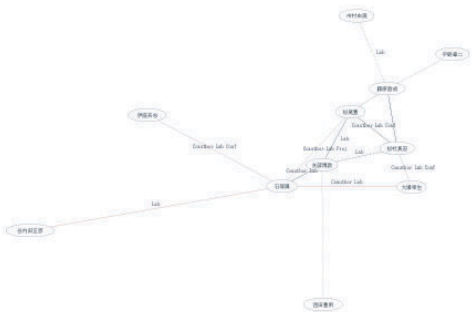

図 3: エゴセントリックネットワーク

参考文献

[1] S. Brin and L. Page. The anatomy of a large-scale hypertextual web search engine. In *Proc. 7th WWW Conf.*, 1998.

[2] Foaf: the 'friend of a friend' vocabulary. http://xmlns.com/foaf/0.1/.

[3] H. Kautz, B. Selman, and M. Shah. The hidden Web. *AI magazine*, Vol. 18, No. 2, pp. 27–35, 1997.

[4] Jon M. Kleinberg. Authoritative sources in a hyperlinked environment. *Proc. ACM-SIAM Symposium on Discrete Algorithms*, pp. 668–677, 1998.

[5] S. R. Kumar, P. Raghavan, S. Rajagopalan, and A Tokins. Trawling the web for emerging cyber communities. In *Proc. 8th WWW Conf.*, 1999.

[6] 村田剛志. 参照の共起性に基づく web コミュニティの発見. 人工知能学会誌, Vol. 16, No. 3, pp. 316–323, 2001.

[7] 中島, 橋田, 森, 伊東, 本村, 車谷, 山本, 和泉, 野田. 情報インフラに基づくグラウンディングとその応用 – サイバーアシストプロジェクトの概要 –. コンピュータソフトウェア, Vol. 18, No. 4, pp. 48–56, 2001.

[8] J. R. Quinlan. *C4.5: Programs for Machine Learning*. Morgan Kaufmann, California, 1993.

[9] Davood Rafiei and Alberto O. Mendelzon. What is this page known for? computing web page reputations. In *Proc. 9th WWW Conf.*, 2000.

[10] 安田雪. ネットワーク分析. 新曜社, 東京, 1997.

[11] 西村, 橋田, 中島. イベント空間支援プロジェクト. 人工知能学会全国大会, No. 3E1-01, 2003.

[12] ユビキタスネットワーク技術の将来展望に関する調査研究会報告書. 総務省, 2001.