オンライン ISSN 1347-4448
印刷版 ISSN 1348-5504

赤門マネジメント・レビュー 4 巻 2 号 (2005 年 2 月)

〔研 究 会 報 告〕社会ネットワーク研究会　2004 年 2 月 6 日

Web 上の情報からの人間関係ネットワークの抽出

友部　博教
名古屋大学情報科学研究科
E-mail: tomobe@nagao.nuie.nagoya-u.ac.jp

1. はじめに

現実世界においても、または Web 上のオンラインの世界においても多種多様なコミュニティが存在している。そのコミュニティは、大学の一研究室のように数名から十数名ほどで成り立つ小さいものから、何千人という会員を抱える学会などの学術組織まで、様々である。そのコミュニティにおいて、人間関係は所属するメンバーを特徴づけるのに大きな役割を持っている。たとえば、コミュニティの中心人物ならば彼の発言は他のメンバーに大きな影響を与えるものになるだろう。またコミュニティがいくつかのクラスタに分かれていれば、それが同じ興味を持つ集まりであることを知ることができる。人間関係を知っていればコミュニティにおけるコミュニケーションも活発にすることができる。たとえば、コミュニティにおける友人の重要度について知ることができる。また、あまり面識のない人物と出会ったとき、自分とはどのような人間関係でつながっているのか知ることができる。

本稿では、学会における人間関係を自動的に抽出する手法について提案する。この方法は、従来の社会学においてのように多くの質問から人間関係を抽出するのではなく、Web 上の情報を探索することによって人間関係を抽出する。人間関係を考慮することによって、我々は人間関係ネットワークを構築する。この人間関係ネットワークにおいて、ノードはコミュニティ内のメンバーを表し、エッジは二人の人間の人間関係を表すことになる。また、エッジに付加されたラベルは、人間関係の種類を表している。このネットワークによって、二人の間の関係の情報を知ることができる。

2. ノードとエッジの抽出

2.1. 基本的なアルゴリズム

人間関係ネットワークを構成するメンバーはあらかじめ決められているとする。つまり、ネットワークのノードは所与である。たとえば、2003 年度の人工知能学会全国大会（以下、

©2005 Global Business Research Center
www.gbrc.jp

JSAI2003 と表記）などの学会の参加者の氏名は、開催に先立って公開されている。また特定の学会誌の過去の論文の著者リスト、情報系の研究者のリストなどを入手すれば、特定の学会や分野における研究者の氏名リストを入手することが可能である。後述するように、同姓同名の問題に対処するために、参加者の氏名に加え所属情報も得る。また、これらの情報は、Web 上に公開されている過去の学会のプログラムから獲得することができる。

次に、ノード間にエッジを付与する処理を行う。

“松尾豊　石塚満”

と入力する（両者は AND の関係である)。ここでヒット件数が多ければ、氏名が共起する傾向が強く、関係が強いであろうことが推測される。本手法では基本的に、Web 文書における氏名の共起の強さによって関係の強さを推測する。

2.2. 共起の強さ

氏名の共起の強さを知るために、両者の名前の AND をとりヒット件数を得ることは有用である。しかし、それを単純に関係の強さの推測値とするのは問題がある。

共起の強さを測るために、共起頻度以外にもさまざまな指標がある。集合の類似度、重なり具合を表す指標として、下記のようにさまざまなものが提案されている (Rasmussen, 1992)。

ここでは、氏名「X」と氏名「Y」の単独でのヒット件数をそれぞれ $|X|$、$|Y|$、AND をとったとき、OR をとったときのヒット件数をそれぞれ $|X \cap Y|$、$|X \cup Y|$と表記する。

共起頻度	$F(X,Y)=\lvert X \cap Y \rvert$	(1)
相互情報量	$-\log \frac{N\lvert X \cap Y \rvert}{\lvert X \rvert \lvert Y \rvert}$	(2)
Dice 係数	$\frac{\lvert X \cap Y \rvert}{\lvert X \rvert + \lvert Y \rvert}$	(3)
Jaccard 係数	$\frac{\lvert X \cap Y \rvert}{\lvert X \cup Y \rvert}$	(4)
Simpson 係数	$\frac{\lvert X \cap Y \rvert}{\min(\lvert X \rvert, \lvert Y \rvert)}$	(5)
Cosine 係数	$\frac{\lvert X \cap Y \rvert}{\sqrt{\lvert X \rvert \lvert Y \rvert}}$	(6)

ここでは、次式で表される閾値付き Simpson 係数を用いた。

$$R(X,Y)=\begin{cases}\frac{|X\cap Y|}{\min(|X|,|Y|)} & if\ |X|>k\ \&\ |Y|>k,\\ 0 & otherwise.\end{cases}$$

JSAI2003 の場合、k=30 とした。つまり単独でのヒット件数が 30 件以下の人はエッジが張られない。（孤立ノードになるので、ネットワークから除外する。）$R(X, Y)$ はネットワークを構築する際に、閾値より高ければエッジを張り、そうでなければエッジを張らないという基準を用いる。また、エッジの長さとして用いることもできる。

2.3. 同姓同名の問題

現実世界と同様に Web 上には同姓同名の人が多数存在し、対象となる人物以外の人物も検索にヒットするために、単独でのヒット件数が多くなることがある。この場合、二つのノードの関係の強さ $R(X, Y)$ を正確に得ることが不可能になるため、所属情報を用いた絞込みを行う。例えば、対象となる人物が産業総合研究所の松尾豊氏の場合

“松尾豊産業技術総合研究所”

というクエリを用い検索する。JSAI2003 の場合には、学会のプログラムに記載されている所属情報を用いた。

2.4. エッジラベルの抽出

社会的関係の種類として、本論文では研究分野に特有の次のようなクラスを定める。これを、エッジのラベルとしてネットワークに付与する。

共著関係　共著の論文がある関係

同研究室関係　同じ研究室や研究所のメンバーなど所属が同じである（あった）関係

同プロジェクト関係　同じプロジェクトや委員会など、組織をまたがる同グループに所属している（いた）関係

同発表関係　同じ研究会や国際会議で発表する（した）関係

以後、誤解のない範囲で、「共著」「研究室」「プロジェクト」「発表」と略記する。関係のクラスは複数取りうる。さて、このような関係を抽出するために、まず検索エンジンに

“X AND Y”

表 1　ページの特徴量

属性	説明	値
NumCo	二人の氏名の共起回数	zero, one, or more than one
SameLine	二人の氏名が同じ行に一度以上出現するか	yes, or no
Rel	関係の強さが閾値以上か	yes, or no
FreqX	X の出現頻度	zero, one, or more than one
FreqY	Y の出現頻度	zero, one, or more than one
GroTitle	タイトルに語群 (A-F) が出現するか	yes or no (それぞれの語群に対して)
GroFFive	最初の 5 行に語群 (A-F) が出現するか	yes, or no (それぞれの語群に対して)

表 2　属性に利用する語群の例

Group	Words
A	出版 (publication), 論文 (papers), 発表 (presentation), 活動 (activities), テーマ (themes), 賞 (awards), 著者 (authors).
B	メンバー (members), 研究室 (lab group), 研究所 (laboratory), 研究機関 (institute), チーム (team).
C	プロジェクト (project), 委員会 (commitee)
D	ワークショップ (workshop), 会議 (conference), セミナー (seminar), ミーティング (meeting), スポンサー (sponsor), シンポジウム(symposium).
E	学会 (society), 団体 (association), プログラム (program), 国立 (national), ジャーナル (journal), セッション (session)
F	教授 (professor), 専攻 (major), 大学院生 (graduate student), 講義 (lecturer).

をクエリとして入力し、上位 5 ページを取得する次に、それぞれのページから表 1 にあるような属性の値を抽出する。属性 GroTitle、GroFFive 属性は、そのページが何に関するページであるかを判断するためのものであり、別に定義した語群を用い、語群 A がタイトルに出現するかどうか (GroTitle (A) 属性)、語群 B が最初の 5 行に出現するか (GroFFive (B) 属性) などを表す。各語群は手動で設定する方法もあるが、できるだけ自動化するために、あらかじめ正解クラスの付与されたページを用い、各クラスごとに TF-IDF 値の上位語を語群とし

表 3　獲得した判別ルールの例

クラス	判別ルール
Coauthor	SameLine = yes
Lab	(NumCo = more than one & GroFFive(F) = no) or (Rel = yes & GroTitle(E) = no & GroFFive(C) = no) or (GroTitle(A) = yes & GroFFive(C) = no & GroFFive(F) = yes) or (GroTitle(E) = no & GorFFive(B) = yes & GroFFive(C) = no) or (GroTitle(E) = no & GroFFive(B) = yes & GroFFive(E) = yes) or (GroTitle(E) = no & GroFFive(B) = yes & GroFFive(F) = yes)
Proj	(FreqX = one & GroTitle(B) = yes) or (GroTitle(C) = yes) or (GroTitle(C) = no & GroFFive(C) = yes & GroFFIve(D) = no & GroFFive(E) = no) or (Rel = no & FreqX = one & GroTitle(B) = yes)
Conf	(FreqY = more than one & GroTitle(D) = yes) or (GroTitle(F) = yes & GroFFive(D) = yes) or (NumCo = zero & GroTitle(F) = no & GroFFive(B) = no & GroFFive(E) = no & GroFFIve(F) = yes) or (NumCO = zero & GroTitle(C) = no & GroFFive(D) = yes & GroFFIve(E) = no) or (NumCO = zero & GroTitle(C) = no & GroTTl(D) = no & GroFFive(D) = yes)

ている。

これらの属性からクラスを判別するルールが得られればよい。これは、属性からクラスを予測するルールを学習する問題となる。そこで本研究では、C4.5 (Quinlan, 1993) を用いて判別ルールを生成する。ランダムに抽出した 275 ページを人手で正解クラスを付与し、これを訓練例として用いた。獲得したルールを表 3 に示す。

3. 人間関係ネットワークの評価

3.1. ラベルの精度

表 4 は判別されたラベルの適合率と再現率を表す。評価には、人間関係ネットワーク上のランダムに選んだ 200 エッジを人手でクラス分けしたものを正解例として求めた。共著ラベルと発表ラベルは他のクラスに比べて高い適合率を得た。表 3 をみればわかるように、この

表 4　ラベルの適合率と再現率

ラベル	適合率	再現率
共著	91.8% (90/98)	97.8% (90/92)
研究室	70.9% (73/103)	86.9% (73/84)
プロジェクト	74.4% (67/90)	91.8% (67/73)
発表	89.7% (87/97)	67.4% (87/129)

二つのクラスは非常に簡単なルールを持っている。また、発表ラベルの再現率が低くなっている。これは、コミュニティのメンバーが参加することのできる研究会が多く存在し、その研究会ごとに Web ページの特徴が異なっているためであると考えられる。

3.2. アンケートによる評価

JSAI2003 の後、人間関係ネットワークに関するアンケート調査を行った。調査対象者は、JSAI2003 に参加登録した人の中から 141 人全員を対象者とした。82 名から回答を得、回収率は 58%であった。一人あたり各 20 人の相手との関係を質問した。質問は、一人の相手あたり各 15 問で「同じ研究室や部署など 30 人規模の組織に同時期に所属していたか」や「研究発表を聞いたり論文を読んだことがある」といった内容である。

表 5 に、JSAI2003 で表示したネットワークのエッジラベルに対して、アンケートから得た回答を正解とした適合率および再現率を示す。また、表 6 は、抽出した全関係に対する適合および再現率であり、ネットワーク中にエッジラベルとして表示していないものも含む。言いかえれば、関係の強さ $R(X, Y)$ が閾値以下のノードペアに対するラベルも含んでいる。

表 4 と比較して、表 5、表 6 は、適合率、再現率ともに低い値になっている。この理由として、回答者が共著やプロジェクトの関係を忘れている、記述もれしているなどの可能性があることや、プロジェクトの定義としてより広いものを想定しており、再現率が低くなっているといった、アンケートの回答に関する問題がある。しかし、最も大きな原因、特に再現率が低いことに対する原因として考えられるのは、この手法は、すべての Web ページを網羅的に分析しているわけではないことや、そもそも、すべての情報が Web 上にあるわけではないということである。そもそも、Web 上にない情報から関係を把握することはできないので、本手法のアプローチが再現率に対して限界があることは明らかである。しかし、関係

表 5　JSAI2003 ネットワークにおけるエッジラベルのアンケートによる評価

クラス	適合率	再現率
共著	89.0% (81/91)	32.1% (81/252)
研究室	78.3% (72/92)	18.7% (72/385)
プロジェクト	50.0% (9/18)	3.0% (9/300)
発表	79.5% (35/44)	6.5% (35/538)

表 6　抽出された全エッジラベルのアンケートによる評価

クラス	適合率	再現率
共著	78.5% (135/172)	53.6% (135/252)
研究室	55.6% (109/198)	28.3% (109/385)
プロジェクト	20.3% (60/296)	20.0% (60/300)
発表	39.9% (222/556)	41.3% (222/538)

の強さと併せて用いることで、表 5 で示したように 80%程度の適合率で共著、研究室、発表などの関係を抽出できるということは、学会におけるコミュニケーション支援という目的には有用であろう。また、本論文で用いた方法はシンプルなものであり、属性の取り方や学習の方法を工夫することで、さらに高い精度を得ることも可能であると考えられる。

4. 関連研究

Referral Web (Kautz, Selman & Shah, 1997) は Web 上で対象人物へのつながりを発見するプロジェクトである。しかし、我々の手法ではまず最初にコミュニティのメンバーのリストを抽出し、そこから人間関係ネットワークを構築するという点で異なる。また関係の種類としてラベルも付加している。

村田らは検索エンジンで検索された件数を用いて Web ページ間の関係を発見する手法を提案した (Murata, 2001)。この手法でも検索件数を用いているが、さらにラベル判別を行い、関係の種類にかんしても言及している点で異なる。

Dan Brickley や Libby Miller は FOAF (Friend of A Friend) という RDF vocabulary を開発し、それを用いて人間関係ネットワークを作成した (Brickley & Miller, 2004; Dumbillm, 2002)。Foaf はそれぞれのユーザが自分の Web サーバ上に Foaf ファイルを作成し URL を共有する。このシステムではユーザは自分自身の人間関係（誰と知り合いか）を記述しなければならない。しかし、このシステムでは自動的にユーザの人間関係を抽出することができる。

5. おわりに

コミュニティにおける人間関係は、そのコミュニティをより緊密にするためには重要な情報となる。本論文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いいのか、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、活発な学会において人間関係ネットワークはどういう特徴を持っているのかなど、今後、様々な研究が可能であると考えられる。

参考文献

Brickley, D. & Miller, L. (2004). *Foaf: The 'friend of a friend' vocabulary*. Retrieved November 2003, from http://xmlns.com/foaf/0.1/

Kautz, H., Selman, B., & Shah, M. (1997). Referral web: Combining social networks and collaborative filtering. *Communications of the ACM, 30*, 63-65.

Murata, T. (2001, May). Finding related web pages based on connectivity information from a search engine. *Proceedings of The Tenth International World Wide Web Conference*, Hong Kong.

Quinlan, J. R. (1993). *C4.5: Programs for machine learning*. San Mateo, CA: Morgan Kaufmann.

Rasmussen, E. (1992). Clustering algorithms. In W. B. Frakes & R. Baeza-Yates (Eds.), *Information retrieval: Data structures & algorithms* (pp. 419-442). Upper Saddle River, NJ: Prentice Hall.

Dumbillm, E. (2002). *XML watch: Finding friends with xml and rdf*. Retrieved November 2003, from http://www-106.ibm.com/developerworks/xml/library/x-foaf.html

赤門マネジメント・レビュー編集委員会

編集長　　新宅　純二郎

編集委員　阿部 誠　粕谷 誠　片平 秀貴　高橋 伸夫　藤本 隆宏

編集担当　西田 麻希

赤門マネジメント・レビュー　4 巻 2 号 2005 年 2 月 25 日発行

編集　東京大学大学院経済学研究科 ABAS/AMR 編集委員会

発行　特定非営利活動法人グローバルビジネスリサーチセンター

理事長　片平 秀貴

東京都千代田区丸の内

http://www.gbrc.jp