1G1-1

# インターネットQAサイトにおけるリンク予測

Link Prediction for Question-Answering Bulletin Boards

森保さき子　　村田剛志
Sakiko Moriyasu　Tsuyoshi Murata

東京工業大学 大学院情報理工学研究科 計算工学専攻
Department of Computer Science, Graduate School of Information Science and Engineering, Tokyo Institute of Technology

Many people use Question-Answering Bulletin Boards(QABB) recently.QABB is a system on the Internet that allows users to answer submitted questions. Communications among users on QABB can be regarded as a social network.Predicting new relations in this network is important for encouraging communications among users. We propose three new weighted graph proximity mesures called weighted Commmon neighbors, weighted Adamic/Adar and weighted Preferential attachment in this paper. These measures are based not only on the proximity of nodes but also on the weights of links in a network. We perform experiments of link prediction with the data of Yahoo! Chiebukuro. Experimental results show that our method is better than previous ones especially for dense networks.

## 1. はじめに

近年インターネット上には人間同士がコミュニケーションをとる場が多く存在し, QA サイトもその一つである.QA サイトとはある人が投稿した質問に対して他者が回答を寄せるシステムであり, そこでの回答者のつながりはソーシャルネットワークと捉えることが出来る. このようなソーシャルネットワークにおいて将来出来るであろう新たな人間関係（リンク）を予測することは, コミュニケーションの促進につながり重要である.

本論文ではネットワークのノード間の類似度を測る従来手法を改良することによって, 新たにリンクの重みを考慮したリンク予測の手法を提案した.Yahoo!知恵袋のデータから構成したネットワークに手法を適用する実験を行い, それぞれどの程度正確にリンク予測できるか比較した. その結果, 提案手法は特に密なネットワークに対して, 従来手法よりも精度が向上することがわかった.

## 2. リンク予測の手法

[Nowell 03] では, 物理学の e-Print arXiv における 5 つの分野について, 論文の著者をノード, 共著関係をリンクとして表したネットワークを用いて将来の共著関係を予測している. 具体的には, ある時刻 t におけるネットワークにリンク予測の手法を適用しネットワークの全てのノードに対して類似度を測ることで, t～t'(t <t') の期間中にネットワークに付け加えられるリンク (新たな共著関係) を予測する. ノード $x, y$ の類似度を $score(x, y)$ と表し, $score$ が大きいノードのペアほど将来新たにリンクが張られやすいと考える.Nowell らは $score$ の計算方法として Common neighbors, Adamic/Adar, Preferential attachment の 3 つを定義しており, それぞれ以下の式で表される. ここで $\Gamma(x)$ はノード $x$ の隣接点集合を表す.

- Common neighbors
  $score(x, y) = |\Gamma(x) \cap \Gamma(y)|$
- Adamic/Adar
  $$score(x, y) = \sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{1}{\log |\Gamma(z)|}$$
- Preferential attachment
  $score(x, y) = |\Gamma(x)| \times |\Gamma(y)|$

連絡先: 森保さき子, 東京工業大学, sakiko@ai.cs.titech.ac.jp

## 3. 重みを考慮した類似度

Nowell らは重みのないネットワークに対して上記の手法を適用し, リンク予測を行っている. しかし実際に存在するネットワークにはノード間の関係が密な部分とそうでない部分が存在し, ネットワークのリンクに重みをつけることによってそれを表現できる. 本論文では, この重みを考慮することでより正確なリンク予測が可能になると考え, Nowell らの手法を改良し, 重みつき Common neighbors, 重みつき Adamic/Adar, 重みつき Preferential attachment を以下のように定義した. ここで $w(x, y)$ はノード $x, y$ 間のリンクの重みを表す.

- 重みつき Common neighbors
  $$score(x, y) = \sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{w(x, z) + w(y, z)}{2}$$
- 重みつき Adamic/Adar
  $$score(x, y) = \sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{w(x, z) + w(y, z)}{2} \times \frac{1}{\log(\sum_{z' \in \Gamma(z)} w(z', z))}$$
- 重みつき Preferential attachment
  $$score(x, y) = \sum_{x' \in \Gamma(x)} w(x', x) \times \sum_{y' \in \Gamma(y)} w(y', y)$$

図 1 は Common neighbors の例である. 図中の数はリンクの重みであり, 重み 2 以上のリンクを太線で示している.

この場合, Nowell らの重みなし手法では $score(x, y)$ は 2 となるが, 重みつき手法では $\frac{(2+1)}{2} + \frac{(1+1)}{2}$ と計算される. リンクの重みは大きいほど互いの重要度が高いことを表している. 重みつき手法ではスコアとして $z \in \Gamma(x) \cap \Gamma(y)$(斜線) とノード $x, y$ の間に張られているリンクの重みの平均をとっており, これによって $x, y$ から見た $z$ の重要度を数値化している.

同様に Adamic/Adar, Preferential attachment の例も図 2, 図 3 に示した.Adamic/Adar では重みなし手法の場合 $\text{score}(x, y) = \frac{1}{\log 4} + \frac{1}{\log 3}$, 重みつき手法の場合 $\text{score}(x, y) = \frac{1.5}{\log 5} + \frac{1}{\log 4}$ となり, Preferential attachment では重みなし手法

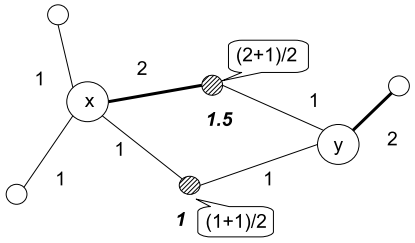


図 1: 重みつき Common neighbors の例

の場合 score$(x, y) = 4 \times 3$, 重みつき手法の場合 score$(x, y) = 5 \times 4$ とそれぞれ計算される.

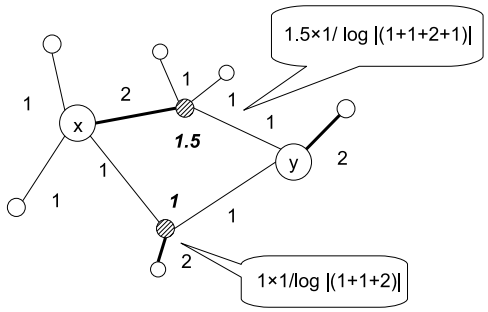


図 2: 重みつき Adamic/Adar の例

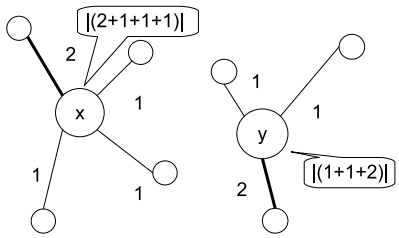


図 3: 重みつき Preferential attachment の例

## 4. 実験

提案手法の精度を確かめるため, Yahoo! JAPAN 提供の Yahoo!知恵袋[*1]研究用デモデータを用いて実験を行った.

### 4.1 実験データ

Yahoo!知恵袋は Yahoo! JAPAN によって運営されている QA サイトである. データには知恵袋に投稿された質問と質問に寄せられた回答が含まれており, それぞれ質問 (回答) 番号, 本文, 質問 (回答) 者 ID[*2], 投稿された日時などから構成されている. 本論文ではこのうち回答者 ID, 日時を用いた. また知恵袋では, 内容によって質問がカテゴリ別に分類されている.

### 4.2 実験手順

実験は以下の手順で行った.

1. 学習期間を 2005 年 9 月 1 日〜15 日, テスト期間を 2005 年 9 月 16 日〜30 日とし, 学習期間のネットワーク $N_{training}$ とテスト期間のネットワーク $N_{test}$ を構成する.

2. $N_{training}$ に含まれる全てのノードのペア $< u, v >$ に対してリンク予測の手法を適用し, 類似度 $score(u, v)$ を計算する.

3. スコアの高いペアほど新たにリンクが出来る可能性が高いと考え, スコアの高いものから順に $|E_{new}|$(学習期間に存在せずテスト期間に新たに出現したリンク数) 個のペアを選ぶ.

4. $N_{test}$ と比較し, 選んだペア間に実際にリンクが出来ていれば予測が正しいと考え, それぞれの手法の精度を算出する. 正解したリンク数を n とすると, 精度 (%)$=\frac{n}{|E_{new}|} \times 100$ と計算した.

ネットワークはノードを Yahoo!知恵袋の回答者とし, ある 2 人の回答者が同じ質問に回答していればノード間にリンクを加えることで構成した. また回答者 2 人が共に回答している質問数でリンクに重みをつけた.

ネットワークはリンクだけでなくノードの増減によっても変化する. 例えば学習期間に存在するノードがテスト期間に存在しないと, その手法によって正しくリンク予測できているかどうか評価できない. そこで学習期間とテスト期間の両方に存在する回答者集合を Core として抽出し, その間に出来るリンクのみを対象として分析を行った. 実験の流れを図 4 に示す.

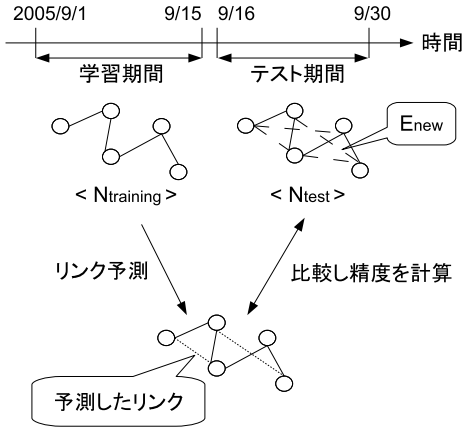


図 4: 実験の流れ

### 4.3 実験結果

Yahoo!知恵袋の 13 のカテゴリについてそれぞれ構成したネットワークの詳細を表 1 に示した.$|E_{old}|$ は学習期間に存在するリンク数である. またこれらに対してリンク予測を行った時の各手法の精度が表 2 である. それぞれ Cn は Common neighbors, AA は Adamic/Adar, Pa は Preferential attachment, Rd は Random を表しており, 上段の数値は重みつき手法, 下段は重

*1 http://chiebukuro.yahoo.co.jp/

*2 質問 (回答) 者 ID は Yahoo!知恵袋で用いられている Yahoo! JAPAN ID そのものではなく, 論文執筆者を含め第三者には解読不可能な文字列に変換されている. そのため投稿者のプロフィール属性は, 論文執筆者を含めた全ての第三者に対して伏せられ保護されている.

みなし手法による精度である.Random は学習期間にリンクが張られていないノード同士をランダムにペアにしたもので, 他との比較に用いている. また各カテゴリは回答者 1 人当たりの次数 (リンク数/ノード数) の降順, つまりネットワークが密な順に並べてある.

表 1: カテゴリ別ネットワーク

| | 学習期間 | | Core | | |
|---|---|---|---|---|---|
| | ノード | リンク | ノード | $\|E_{old}\|$ | $\|E_{new}\|$ |
| Yahoo! JAPAN | 5820 | 94449 | 1290 | 43525 | 44963 |
| ニュース | 4992 | 76070 | 862 | 26889 | 22055 |
| 健康 | 9991 | 149230 | 2351 | 66522 | 67970 |
| 子育て | 5804 | 76197 | 1133 | 29243 | 30572 |
| マナー | 2782 | 29535 | 414 | 8183 | 7733 |
| スポーツ | 4789 | 49937 | 1102 | 21459 | 22699 |
| エンターテイメント | 7454 | 68538 | 1411 | 22977 | 29966 |
| 暮らし | 5409 | 40026 | 985 | 13859 | 14736 |
| 教養 | 4568 | 28486 | 813 | 8442 | 8963 |
| 地域 | 3109 | 17327 | 470 | 4562 | 4575 |
| ビジネス | 2103 | 10533 | 278 | 2198 | 2658 |
| インターネット | 3198 | 13573 | 575 | 5111 | 5106 |
| 職業 | 2179 | 7811 | 206 | 807 | 893 |

表 2 より, Common neighbors, Adamic/Adar, Preferential attachment の 3 手法はいずれも Random の 5〜12 倍の精度を示していることがわかる.

## 5. 考察

Nowell らの実験では Core の著者数は 486〜1790, $|E_{old}|$ は 519〜6654, $|E_{new}|$ は 400〜5751 である. 表 1 より, 本論文の Yahoo!知恵袋の社会ネットワークのリンク数はこの 10 倍程度の規模であることがわかる.Yahoo!知恵袋の社会ネットワークは論文の著者のネットワークよりもオープンであり, 参加しているユーザ数も多い. このように全く異なる社会ネットワークにおいても, ネットワーク構造に基づいたリンク予測手法が有効であることが示せた.

また Web 上の社会ネットワークにおいては個人情報の悪用を避けるため, ユーザの属性（年齢, 性別, 職業等）は伏せられたり, 偽られたりしている場合も多いが, 本論文で述べたネットワーク構造の近接性を用いた手法はそれらの属性を全く使っていない. このことからも, 本論文のリンク予測手法は Web 上の社会ネットワークに対して有効であると考えられる.

以下で実験結果に基づいて各手法の特徴を分析する. 表 1 に示したネットワークのノード次数別人数分布と次数の最大値が表 3 である. それぞれを比較しやすいように, カテゴリごとに次数の最大値を 100 %としており, 数字は次数 $x$ %までのノード数を表している. 例えば「健康・美容とファッション」は次数の少ないノードが大半であることがわかる.

### 5.1 重みなし手法の考察

- **3 手法は密なネットワークに対して精度が良い**
  Common neighbors は 14〜29 %, Adamic/Adar は 15〜29 %, Preferential attachment は 12〜27 %と, どの手

表 2: カテゴリ別精度 (%)

| | Cn | AA | Pa | Rd |
|---|---|---|---|---|
| Yahoo! JAPAN | 32.0 | 32.2 | 24.7 | |
| | 29.5 | 29.9 | 24.5 | 2.8 |
| ニュース・政治・国際情勢 | 25.2 | 25.4 | 25.9 | |
| | 23.5 | 23.8 | 25.2 | 3.1 |
| 健康・美容とファッション | 17.4 | 16.9 | 17.1 | |
| | 15.7 | 16.0 | 16.6 | 1.3 |
| 子育てと学校 | 22.9 | 23.0 | 22.0 | |
| | 20.5 | 22.3 | 19.4 | 2.4 |
| マナー・冠婚葬祭 | 30.2 | 30.3 | 27.6 | |
| | 29.2 | 29.4 | 27.5 | 5.3 |
| スポーツ・アウトドア・車 | 25.4 | 25.6 | 15.9 | |
| | 23.2 | 24.8 | 16.2 | 2.1 |
| エンターテイメントと趣味 | 16.1 | 16.1 | 14.6 | |
| | 15.2 | 15.3 | 14.5 | 1.6 |
| 暮らしと生活ガイド | 18.7 | 19.2 | 18.9 | |
| | 18.2 | 18.3 | 18.7 | 1.5 |
| 教養と学問・サイエンス | 15.9 | 16.4 | 12.3 | |
| | 15.8 | 16.1 | 12.3 | 1.4 |
| 地域・旅行・お出かけ | 22.0 | 22.0 | 15.2 | |
| | 20.1 | 20.5 | 16.0 | 2.3 |
| ビジネス・経済とお金 | 26.3 | 27.6 | 19.0 | |
| | 26.3 | 26.9 | 19.6 | 3.6 |
| インターネット・PC と家電 | 18.9 | 19.4 | 17.9 | |
| | 18.6 | 19.2 | 17.5 | 1.5 |
| 職業とキャリア | 14.9 | 16.9 | 15.0 | |
| | 14.5 | 16.9 | 16.6 | 2.2 |
| 平均 | 22.0 | 22.4 | 18.9 | |
| | 20.8 | 21.5 | 18.9 | 2.4 |

表 3: カテゴリ別次数分布と最大次数

| | 次数分布 | | | 最大次数 |
|---|---|---|---|---|
| | 〜10 % | 20〜60 % | 70 %〜 | |
| Yahoo! JAPAN | 3763 | 1963 | 94 | 1070 |
| ニュース | 3492 | 1465 | 35 | 1764 |
| 健康 | 9478 | 509 | 4 | 4356 |
| 子育て | 3939 | 1819 | 46 | 986 |
| マナー | 1156 | 1566 | 60 | 591 |
| スポーツ | 2539 | 2147 | 103 | 483 |
| エンターテイメント | 5392 | 2020 | 42 | 749 |
| 暮らし | 4716 | 687 | 6 | 1195 |
| 教養 | 2602 | 1910 | 56 | 409 |
| 地域 | 2010 | 1085 | 14 | 454 |
| ビジネス | 915 | 1165 | 23 | 254 |
| インターネット | 2869 | 323 | 6 | 932 |
| 職業 | 1196 | 973 | 10 | 322 |

法もカテゴリによってだいぶ差がある. 特に精度 20 %未満の「健康・美容とファッション」「エンターテイメントと趣味」「暮らしと生活ガイド」「インターネット・PC と家電」「職業とキャリア」に着目すると, 表 3 においていずれも次数上位 (70 %〜) の人数が少ない. 逆に「マナー・冠婚葬祭」「ビジネス・経済とお金」では比較的次数 10 %以下の人数が少なく, 3 手法の精度がほぼ 25 %以上である. このことからネットワーク内に次数の低いノードが多く存在するよりも, 高いノードが多く存在する方が精度が上がるのではないかと推測できる.

次数の低いノードばかりだとネットワークが疎になり, ノード $x$ の隣接点 $\Gamma(x)$ の個数は少ない. すると Common neighbors, Adamic/Adar においてノード $x, y$ に共通な隣接点の個数 $|\Gamma(x) \cap \Gamma(y)|$ も少なくなるため, どのペアのスコアも低い値になってしまい良く予測できないと考えられる. 次数の高いノードが多い方が, $|\Gamma(x) \cap \Gamma(y)|$ の値に差が出て精度が上がる.

- **Adamic/Adar は Common neighbors よりも精度が良い**
  全てのカテゴリにおいて Adamic/Adar の方が Common neighbors より精度が良い. Adamic/Adar では $z \in \Gamma(x) \cap \Gamma(y)$ の次数を考慮することでより詳しく計算しており, 張られているリンク数の少ない $z$ ほど $z$ から見た $x, y$ の重要度が大きくなる.

  また Nowell らの実験結果でも, Adamic/Adar は Common neighbors より良い精度を出している場合が多いことから, Adamic/Adar はデータによらず比較的高精度な予測ができる手法であると言える.

- **Preferential attachment はノード次数の差が大きいネットワークに対して精度が良い**
  Preferential attachment は Common neighbors, Adamic/Adar と比べて精度が悪い場合が多い. Preferential attachment は次数の高いノードほど新たにリンクが張られやすいという考え方に基づいているため, 次数の高いノードを含むペアほどスコアが高くなるはずである. しかし図 5 の場合, $y1$ の次数が最も高いにも関わらず $x2, y2$ のペアのスコアの方が高くなってしまっている. このようにネットワーク内のノード次数の高低差があまりなく, どのノードの次数も一様に低いと, スコアの順位が逆転してしまい正しくリンク予測できない可能性がある.

  「ニュース・政治と国際情勢」「健康・美容とファッション」では他 2 手法と比べて精度が落ちていない. これは表 3 よりノードの最大次数が 1764, 4356 と比較的高いためと考えられる. 逆に「教養と学問・サイエンス」「地域・旅行・お出かけ」「ビジネス・経済とお金」などの最大次数の低いカテゴリでは Preferential attachment の精度が他 2 手法よりも落ちていることが多い. またこれらのカテゴリの次数分布を見ると, Preferential attachment の精度差は次数の高い（または低い）ノードの割合によらないことがわかる. よってネットワークにおけるノードの次数の差が Preferential attachment の精度に影響を与えていると推測できる.

### 5.2 重みつき手法の考察

- **Preferential attachment は重みつき手法の方が重みなし手法よりも精度の悪い場合が多い**

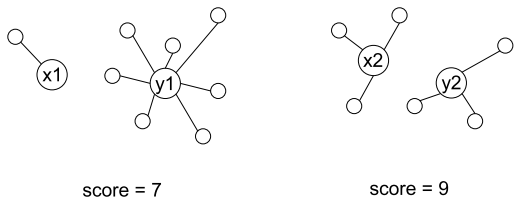


図 5: Preferential attachment : 悪い例

重みなし手法の方が精度が良いのは, Common neighbors, Adamic/Adar は 13 カテゴリ中 0 カテゴリなのに対し, Preferential attachment は 5 カテゴリもある. 図 6 において, $x1$ に張られているリンクの重みが全て 1, $x2$ に張られているリンクの重みが全て 4 だとすると, $\sum_{x1' \in \Gamma(x1)} w(x1', x1) < \sum_{x2' \in \Gamma(x2)} w(x2', x2)$ である. すると $x2$ の方が次数が低いにも関わらず, $x2$ を含むペアのスコアの方が $x1$ を含むペアより高くなる可能性がある.5.1 節と同様に, これは Preferential attachment の本来の考え方に沿っていない. よって重みつき手法の精度が悪くなっていると考えられる.

逆に Common neighbors は, 3 手法の中で重みつけによる精度の上昇幅が最も大きく, 適切な重みつけができていると推測できる.

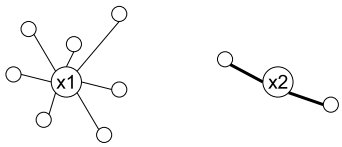


図 6: 重みつき Preferential attachment : 悪い例

## 6. 結論

本論文ではネットワークのノード間の類似度とリンクの重みの両方を考慮する新たなリンク予測の手法を提案した. またインターネット QA サイトに対して実験を行った結果, 提案手法は従来手法よりも精度が向上することを確認した.

今後の課題としては, 学習期間のデータのうち最近のものをより重要視することによって, さらに精度を向上させることが考えられる.

## 7. 謝辞

Yahoo!知恵袋のデータ利用に際し, ヤフー株式会社の岡本真様と, 国立国語研究所の前川喜久雄先生に大変お世話になりました. 深くお礼申し上げます.

## 参考文献

[Nowell 03] Nowell, D. L., Kleinberg, J.: The Link Prediction Problem for Social Networks, Proc of 12th International Conference on Information and Knowledge Management(CIKM), 556-559(2003).