DBSJ Journal, Vol.8, No.1
June 2009

QA コミュニティの成長パターンに基づく回答者への質問推薦

Questions Recommendation Based on Evolution Patterns of a QA Community

甲谷 優♥　川島 晴美♠
藤村 考

Yutaka KABUTOYA　Harumi KAWASHIMA
Ko FUJIMURA

これまでの QA サイトでは，回答者は回答したい質問を自分で見つけてくる必要があった．そこで，回答者に適切な質問者の質問を推薦する手法を提案する．具体的には，ユーザをノード，回答をエッジとして QA サイト上のコミュニケーションをグラフにモデル化したものである QA ネットワークに着目する．その局所構造における新たなエッジのつきやすさに基づき，コミュニケーションをとりそうな 2 者を発見する．本稿では，提案手法に関して説明し，さらにその評価実験の結果について報告する．

In an existing QA site, it is necessary for repliers to retrieve attractive questions for themselves. In this paper, we propose a method to retrieve questions for which each replier is likely to give an answer. In our method, based on the possibility of a new answer in local structure of the "QA network", we discover two users between whom a new answer is likely to be derived. The "QA Network" is such a graph that each node mean a user and each edge mean an answer. In this paper, we describe our method and report experimental evaluation of our method.

1. はじめに

QA サイトとは，あるユーザの自然文で表わされた質問に対して別のユーザが回答することによる，人どうしの知識の共有を目的としたサービスである．

質問に回答が多くつくことは，QA サイトを活性化するにあたって重要である．そこで本研究では，ユーザ u の質問に対し回答しそうなユーザが v であることを予測し，それに基づきユーザ v に対してユーザ u の質問を推薦する手法を提案する．このことにより，回答者にとっては興味のある質問を発見しやすくなり，かつ質問者にとっては適切な回答をする回答者から回答されやすくなると考えられる．

ユーザ u の質問に対し回答しそうなユーザが v であることを予測する手法としてもっとも簡単なものは，過去のユーザ u の質問にユーザ v が回答した頻度を用いる方法が考えられる．しかしこの方法では，一度も回答が発生したことのない 2 ユーザを発見することができず，QA サイトの活性化にはつながらない．

♥ 正会員　日本電信電話株式会社　NTT サイバーソリューション研究所　kabutoya.yutaka@lab.ntt.co.jp
♠ 非会員　日本電信電話株式会社　NTT サイバーソリューション研究所　{kawashima.harumi, fujimura.ko}@lab.ntt.co.jp

そこで本研究ではユーザ u の質問に対し回答しそうなユーザが v であることを予測するために，QA コミュニティをグラフにモデル化したものである QA ネットワークの成長パターンに着目し，そのリンク予測を行う．QA ネットワークの各ノードはユーザ，各エッジは回答（エッジの向きは回答者から質問者）を表す．

特に本研究では，QA ネットワークの局所構造における成長に関する制約に基づいてリンク予測を行う．まず，ある時点 t_0 の QA ネットワークに含まれる 3 ノードからなる部分グラフ構造をすべて抽出しておく．次に，時刻 t_0 から t_1 の間の回答について，抽出した 3 ノード部分グラフ内の場所ごとに発生回数を算出する．その情報を用いて QA ネットワークにおける 3 ノード部分グラフのどこに回答が発生しやすいか評価する．次に，時刻 t_1 における 3 ノード部分グラフを抽出し，それらの情報を用いて時刻 t_1 から t_2 の間に回答が発生しやすそうな 2 ユーザを発見する．

2. 関連研究

2.1 ネットワークモチーフ

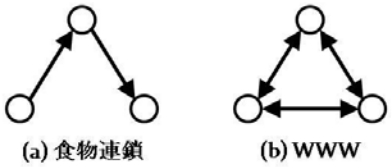

図 1　ネットワークモチーフの例
Fig.1　Examples of network motifs

複雑ネットワークの局所構造を分析するために，Milo ら [1]によって提唱された**ネットワークモチーフ分析**と呼ばれる手法に注目が集まっている．ネットワークモチーフとは，観測された複雑ネットワークが，ノード数・エッジ数・各ノードに入るエッジの数（入次数）と各ノードから出るエッジの数（出次数）等が同一のランダムネットワークと比較して頻繁に出現する部分グラフのパターンを指す．異種のネットワークでは発見されるネットワークモチーフも異なり，それまで発見されていたネットワークの広域特性（たとえばスモールワールド性）では不可能であったネットワークの分類が可能になった．たとえば 3 ノードのネットワークモチーフの場合，食物連鎖では図 1 (a) に示すようなパターンが頻出し，WWW では図 1 (b) に示すようなパターンが頻出することが発見された．

ネットワークモチーフ分析が盛んになった理由の 1 つに計算機性能の大幅な向上がある．なぜならば，ネットワークモチーフ分析には必ず部分グラフ抽出のステップが必要になるが，それには莫大な計算量がかかるからである．たとえば，ノード数 n のグラフから k ノードの部分グラフを抽出するために必要な計算量は，$O(n^k)$ に達する．

しかしネットワークモチーフ分析は，ランダムネットワークと比較することによる部分グラフの各パターンの相対頻度を求める問題であり，部分グラフを全数調査し各パターンの絶対的な頻度を求める必要はない．そこで Wernicke ら [2] は，部分グラフのサンプリングによるネットワークモチーフ分析の高速な計算手法を提案している．このように，ネットワークモチーフ分析の高速化に関する研究は多い．

しかしながら，ネットワークモチーフの応用となる研究は

未だに少ない．高田ら [3] は 3 ノードからなる部分グラフ内の各ノードを部分グラフのパターンとノードの位置から 30 種類に分け，従来のネットワークの隣接行列にそのノードの種類を加味し次元を拡大することにより Wikipedia を対象とした階層的クラスタリングを行った．結果，顕著な効果があったと報告されている．

本提案手法においても，部分グラフ内の 2 ノード間の回答の発生可能性を評価するために，観測されたものとランダムに発生させたものを比較している．この意味で，本研究は部分グラフ内のエッジに着目した一種のネットワークモチーフ分析といえよう．また，回答者への質問推薦に用いるという点で，これまであまりなされてこなかったネットワークモチーフの応用に関する研究と位置づけることができる．

2.2 リンク予測アルゴリズム

「ネットワークの既知の部分が与えられたとき，未知の部分を予測する」リンク予測問題に関する研究は数多い．本研究もその問題に対する取り組みの 1 つであると言える．

Newman [4] は，2 つのノードに共通して隣接するノードの数で 2 ノード間のエッジの存在可能性を評価した．つまり，「友人の友人はやはり友人である可能性が高い」という仮説に基づく手法である．この手法はノードの次数に大きく影響を受ける．一方で，2 ノードの隣接ノードの Jaccard 係数を取る手法も提案されている [5,6]．この手法であれば次数の低い 2 ノードであってもエッジの存在可能性を評価することができる．Adamic ら [7] は共通隣接ノードのそれぞれに次数の逆数で加重する手法を提案した．この手法は「友人の少ない友人を共通に友人としてもつ 2 人は友人である可能性が高い」という仮説に基づく手法である．もっともクラシックな手法としては，グラフ内の注目 2 ノード間のすべてのパスを発見し，それらの数で評価する手法がある [8]．この手法は [4] の一般化であると考えることができる．

ただし，これら「共通の友人」に基づく手法は，「共通の友人」が存在しないようなネットワークに適用するのは不向きである．観測されるグラフのスケールフリー性に着目した手法も存在する [4,9]．この手法は，「友人の多い人間はさらに友人を作りやすい」という仮説に基づく．この手法では，ノード数が増え続けているような未成熟なネットワークに適用するには不向きである．

2.3 QAネットワーク分析

近年，本研究と同様に QA ネットワークを対象とした分析が盛んである．我々は以前，QA ネットワークの成長パターンについて分析を行い，その報告を行った [10]．

Adamic ら [11] は Yahoo Answers を対象として，カテゴリごとにさまざまな手法を用いて多角的に分析を行った．その結果，カテゴリ内で行われているコミュニケーションタイプとして主に「知識共有」「相談」「議論」の3 種類が存在することを示した．

大平ら [12] は，ネットワークの成長パターンについて，Yahoo!知恵袋のデータや，Apache コミュニティ等を比較した．村田ら [13] は，Yahoo!知恵袋の QA ネットワークについて，リンク予測アルゴリズムや，コミュニティ抽出アルゴリズムを適用した結果について報告している．

3. QA ネットワークとその成長

まず QA ネットワークの定義について言及し，さらに時間とともにノード・エッジが増大する QA ネットワークの性質について説明する．

QA サイトでは，1 つの質問に対して複数の回答が付く．質問，回答にはそれぞれ発信者のユニーク ID が付与されている．それらユニーク ID から識別されるユーザをノードとし，回答を回答者から質問者へのエッジとする．そうすると 1 つの質問は，図 2 のように複数の回答者から 1 人の質問にエッジが貼られるようなグラフにモデル化できる．

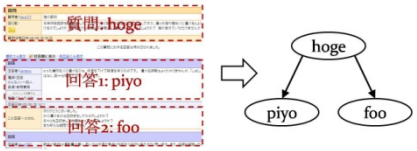

図 2　QA ネットワーク
Fig.2　A thread and its QA network

また，回答には発信時刻が付与されている．ここで，回答 a は時刻 t，質問者 u，回答者 v により一意に決まるものとする．この回答が回答者 v に対応するノード n_v から質問者 u に対応するノード n_u へのエッジ e_{vu} を発生させることにする．このとき，生成時刻が t 以下の回答のノードとエッジのユニオンを時刻 t の QA ネットワークと定義する．このように定義すると，QA ネットワークは時間経過とともにノード・エッジが増大する．

4. 局所構造に基づく回答者への質問推薦

本節では，ユーザ v が，実際に存在する質問の中でどのユーザの質問に回答しそうかを予測する手法を提案する．すなわち，本提案手法は質問者集合 U と，回答者集合 V の 1 人 1 人 v がクエリとして入力されたときの適切な U の順序付き部分集合を出力とする検索のランキングに帰着する．

4.1 仮説

近年注目されているネットワークモチーフ分析の結果から，ネットワークにはその生成段階，成長において何らかの制約条件があり，それに左右された結果特定のパターンの部分グラフが頻出するようになると考えられている．本提案手法では，部分グラフ構造に現れるネットワークの一般的な成長に関する制約を，成長段階のネットワークに含まれる部分グラフから推定し，ネットワークがその制約に沿ってさらに成長を続けると仮定してリンク予測を行う．

具体的には，QA ネットワークに含まれる 3 ノード部分グラフのパターンに注目する．図 3 に，連結である 3 ノード部分グラフの全 13 通りのパターンを示す．

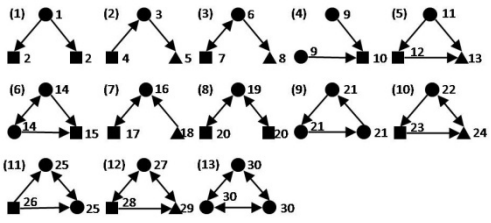

図 3　3 ノード部分グラフのパターンとノードタイプ
Fig.3　Patterns of 3-node subgraphs and node types

さらに，各パターン内のノードは，その位置と対称性を考慮して図 3 の通り 30 種類に分類することができる．30 種類のノードタイプを考慮するのは，高田ら [3] による手法と

同様である．さらに我々は，30 種類のノードタイプ間のエッジが存在する，あるいはエッジが発生しうるような回答発生箇所を考える．たとえば，パターン (1) の部分グラフにおいて，ノードタイプ 1，2 が存在し，回答が発生しうる箇所は 1→2，2→1，2→2 の 3 種類となる．このようにして考えると，3 ノード部分グラフに基づく回答発生箇所の種類は計 54 得られる．

54 の回答発生可能なノードタイプ間のうち，QA ネットワークの成長の制約により回答が発生しやすい箇所と発生しにくい箇所が存在すると考えられる．たとえば，2→1 や 3→4 の箇所では回答は発生しづらいが，2→2 や 4→5 の箇所では回答が発生しやすいという予測が立つ．

4.2 アルゴリズム

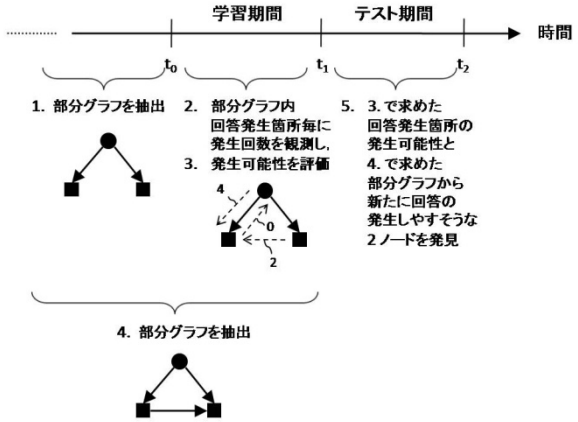

図4　アルゴリズムの概要
Fig. 4　Overview of our algorithm

図 4 に，アルゴリズムの概要を示す．入力としては，時刻 t_0 の QA ネットワーク Q_0 と，時刻 t_1 の QA ネットワーク Q_1 で，出力としては時刻 t_1 から t_2 までに発生する回答$\{a \mid t_1 < t < t_2, n_u, n_v \in Q1\}$になる．出力が $Q_2 - Q_1$ に含まれる新規エッジのみでないことから，通常のリンク予測アルゴリズムとは異なることに注意されたい．

ステップ1:　Q_0 内の部分グラフの抽出

時刻 t_0 のQAネットワーク Q_0 より，3 ノード部分グラフの抽出を行う．ただし，ネットワークモチーフ分析とは異なり各パターンの出現頻度を求めるのではなく，連結な 3 ノードをすべて列挙し，それらがどのパターンをなしているか，そして各ノードが 30 種類のどれかに関する情報を抽出する．

ステップ2:　ノードタイプ間ごとの回答発生回数の算出

時刻 t_0 から時刻 t_1 までに発生した回答の各々が，どの 2 ユーザ間で発生したかに基づき，ステップ 1 で求めた 3 ノード部分グラフに関する情報から 54 種類のノードタイプ間のどの箇所にあたるかをカウントする．ただし，ある回答が，複数の 3 ノード部分グラフに出現するような 2 ユーザ間で発生することはままある．そのような場合は，対応するすべての部分グラフに対し，その個数分対応するノードタイプ間の回答発生回数をカウントする．2 ユーザのうち 1 人でも時刻 t_0 までに出現しないような新規ノードであったり，あるいは 3 ノード部分グラフに含まれないような 2 ユーザのときは，いずれの箇所もカウントしない．

ステップ3:　ノードタイプ間ごとの推薦スコア算出

ステップ 2 にて算出した回答発生箇所ごとの回答発生回数をそのまま 2 ユーザ間の推薦スコアに適用してもよいかもしれない．しかし，3 ノード部分グラフの各パターンの出現頻度には大きな偏りがある．たとえば QA ネットワークには双方向エッジがほとんど存在しないという事実から，3 ノード部分グラフパターン (13) はほとんど存在しない．したがって，必然的に answerFreq(30,30) は低くなってしまう．逆に，構造的にパターン (1) のようなものは数多く存在し，数多くの 3 ノード部分グラフパターン (1) のノードタイプ 1，2 をなしている 2 ユーザ間で回答が発生してしまうとその数だけ answerFreq(1,2) が加算されてしまう．

ネットワークモチーフ分析における 3 ノード部分グラフの頻度も同様の問題が存在する．ネットワークモチーフ分析では，観測された複雑ネットワークと等価なランダムネットワークを生成し，それに出現する 3 ノード部分グラフの出現頻度と比較し正規化することでその問題に対処している．

そこで本提案手法においても，時刻 t_0，t_1 間で発生した回答と等価にランダムに 2 ユーザを選択していき，比較対象として擬似回答を発生させる (以後，これをランダムアンサと呼ぶことにする)．ただし，完全にランダムに 2 ユーザを選択したのでは，ランダムアンサは実際に時刻 t_0，t_1 間で発生した回答と等価にならない．

まず，発生させる擬似回答の回数は，実際のそれと同じでなければならないのは自明である．観測される実際の回答は，スケールフリーネットワークを構築していく．ノードの入次数・出次数は実際の回答発生後と，ランダムアンサ発生後のそれぞれの QA ネットワークで同じでなければならない．つまり，各ユーザの質問回数・回答数は保持される形でランダムに擬似回答を発生させる必要がある．

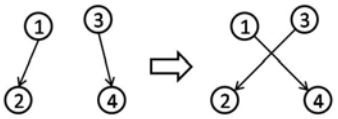

図5　EXPLICIT におけるランダムネットワーク生成
Fig. 5　Generation of random networks in the EXPLICIT algorithm

Wernicke [2] らは，ランダムネットワーク生成によるネットワークモチーフ分析を EXPLICIT と呼んでいる．この手法では，図 5 に示すように，同一グラフ内の 2 つのエッジの始点と終点を交換することにより，ノードの次数・ノード数・エッジ数で等価なランダムネットワークを生成する．

本研究でも，同様の手法でユーザ毎の質問回数・回答数，ユーザ数，総回答数が等価なランダムアンサを複数回生成し，ステップ 2 と同様の手法でそれぞれのノードタイプ間の回答の発生回数を求め，それらの平均 answerFreq(p_i,p_j) と標準偏差 $\sigma(p_i,p_j)$ を求める．このときノードタイプ間の推薦スコア structureScore(p_i,p_j) を観測された answerFreq(p_i,p_j) の answerFreq(p_i,p_j)，$\sigma(p_i,p_j)$ に対する Z 値で与える．

$$\mathrm{structScore}(p_i,p_j) = \frac{\mathrm{answerFreq}(p_i,p_j) - \overline{\mathrm{answerFreq}}(p_i,p_j)}{\sigma(p_i,p_j)} \quad (1)$$

ステップ4:　Q_1 内の部分グラフの抽出

ステップ 1 と同様の手法を用いて，時刻 t_1 の QA ネットワーク Q_1 より 3 ノード部分グラフの抽出を行う．

表2　回答発生箇所別推薦スコア
Table2　Calculated weights of local structure for our recommendation

回答発生箇所		自転車	スイーツ	海外生活	クラシック	手芸	PHP	音楽(教科)	栄養	留学	Java
(1)	1→2	0.220	-0.250	4.431	-0.395	0.006	-0.392	-0.411	-0.099	-0.264	-0.098
	2→1	4.927	-0.303	2.525	3.085	4.652	2.768	1.032	0.846	3.912	0.992
	2→2	1.391	-0.608	-0.433	-0.468	-0.649	-0.200	-1.289	-0.807	-0.031	1.771
(2)	3→4	0.948	2.175	0.820	1.830	2.734	0.585	0.114	-0.637	5.403	0.874
	3→5	1.239	-0.243	3.134	1.048	1.251	-0.216	-0.421	-0.054	-0.275	-0.657
	4→3	1.231	-0.241	0.324	-0.478	-0.501	-0.864	-0.206	-0.404	-0.295	-0.686
	4→5	-0.433	-0.246	4.899	-0.081	-0.366	0.992	0.727	-0.240	-0.325	-0.407
	5→3	−	−	−	−	−	−	−	−	−	−
	5→4	−	−	−	−	−	−	−	−	−	−
(3)	6→7	−	−	−	7.230	−	-0.737	−	−	−	-0.123
	6→8	−	−	−	5.932	−	−	−	−	−	-0.045
	7→6	−	−	−	6.273	−	-0.737	−	−	−	-0.133
	7→8	−	−	−	-0.886	−	-0.095	−	−	−	-0.303
	8→6	−	−	−	−	−	−	−	−	−	−
	8→7	−	−	−	−	−	−	−	−	−	−
(4)	9→9	2.936	1.905	2.010	0.344	3.329	2.301	-0.396	-0.480	4.352	0.563
	9→10	1.047	0.518	4.988	0.904	4.584	3.464	0.115	-0.136	1.370	1.554
	10→9	−	−	−	−	−	−	−	−	−	−
(5)	11→12	6.084	1.279	4.493	-0.873	4.676	2.534	-0.406	-0.331	5.852	-0.035
	11→13	0.480	-0.048	3.865	2.083	-0.149	-0.077	-0.546	0.990	-0.639	-0.427
	12→11	1.134	3.077	3.557	2.028	-0.156	3.043	1.809	-0.273	4.150	0.317
	12→13	0.252	-0.048	3.214	-0.896	-0.262	1.091	-0.504	−	1.197	0.398
	13→11	−	−	−	−	−	−	−	−	−	−
	13→12	−	−	−	−	−	−	−	−	−	−
(6)	14→14	−	−	−	7.256	−	−	-0.128	−	−	−
	14→15	−	−	−	5.590	-0.045	−	−	−	−	-0.289
	15→14	−	−	−	−	−	−	−	−	−	−
(7)	16→17	−	−	−	5.856	−	−	-0.128	−	−	−
	16→18	−	−	−	6.816	−	−	2.005	−	−	2.116
	17→16	−	−	−	5.607	−	-0.737	−	−	−	−
	17→18	−	−	−	-0.795	−	-0.115	0.818	−	−	-0.101
	18→16	−	−	−	−	−	−	−	−	−	−
	18→17	−	−	−	−	−	−	−	−	−	−
(8)	19→20	−	−	−	−	−	−	−	−	−	−
	20→19	−	−	−	−	−	−	−	−	−	−
	20→20	−	−	−	−	−	−	−	−	−	-0.071
(9)	21→21	−	−	−	-0.055	−	−	-0.207	−	−	−
	21←21	−	−	−	−	−	−	−	−	−	−
(10)	22→23	−	−	−	-0.988	−	−	−	−	−	−
	22→24	−	−	−	4.776	−	−	−	−	−	−
	23→22	−	−	−	0.587	−	−	−	−	−	−
	23→24	−	−	−	-0.427	−	−	−	−	−	−
	24→22	−	−	−	−	−	−	−	−	−	−
	24→23	−	−	−	−	−	−	−	−	−	−
(11)	25→25	−	−	−	4.776	−	−	−	−	−	−
	25→26	−	−	−	0.806	−	−	-0.220	−	−	-0.160
	26→25	−	−	−	-0.763	−	−	−	−	−	−

ステップ5:　2 ユーザ間の推薦スコア算出

ステップ 3 にて求めたノードタイプ間ごとの推薦スコアと，ステップ 4 にて求めた Q_1 に含まれる 3 ノード部分グラフから 2 ユーザ間の推薦スコアを算出し，そのスコアをもとに時刻 t_1，t_2 間に発生しそうな回答を予測する．

今，ユーザ n_u の質問の，n_v に対する推薦スコアを求めることとする．ステップ 4 にて列挙した Q_1 に含まれる 3 ノード部分グラフの中で，n_u，n_v をともに含むものに着目し，その数を K とする．そのような 3 ノード部分グラフそれぞれにおける n_u，n_v のノードタイプ (1～30 のいずれかの値) を (p_u^k, p_v^k) $(k = 1,2,\ldots,K)$ とする．このとき n_u，n_v の推薦スコア $\mathrm{score}(n_u,n_v)$ を以下の式で定義する．

$$\mathrm{score}(n_u, n_v) = \sum_k \mathrm{structScore}(p_u^k, p_v^k) \qquad (2)$$

5. 評価実験

5.1 実験環境と手順

ここで，提案手法の有効性を検証するために行った評価実験について報告する．まず実験データとして，教えて！goo[1] における全 309 あるカテゴリのうち，10 種類のカテゴリを選択し，それぞれからデータを取得して評価実験を行った．また，アルゴリズムに用いる 3 点時刻 t_0，t_1，t_2 を，それぞれ 2006 年 1 月，2007 年 1 月，2007 年 8 月と設定した．表 1 に実験データに用いたカテゴリと，各時点それぞれで含まれる質問・回答数を示す．

まずそれぞれのカテゴリについて，時刻 t_0 (= 2006 年 1 月)までにおける QA ネットワーク Q_0 に含まれる 3 ノー

[1] http://oshiete.goo.ne.jp/

ド部分グラフを抽出する（ステップ 1).

表 1　実験データ
Table 1　Experimental data

カテゴリ	t_0		t_1		t_2	
	質問数	回答数	質問数	回答数	質問数	回答数
自転車	1434	4746	2756	9200	3900	12807
スイーツ	2265	7103	3073	9281	3685	10688
海外生活	1720	5079	2874	8586	3145	9382
クラシック	1393	4506	2383	7665	2986	9269
手芸	3116	7128	4259	9450	4910	10681
PHP	2425	5112	4208	8864	5540	11499
音楽(教科)	2236	7495	3356	10974	4044	12955
栄養	2481	7761	3396	10443	3914	11880
留学	1786	5747	2971	9106	3756	11175
Java	4449	9840	5872	13030	6838	15084

次に，ノードタイプ間ごとの回答発生回数を求め（ステップ 2)，ランダムアンサとの比較によってノードタイプ間ごとの推薦スコアを算出する（ステップ 3). 付録の表 2 に，我々の手法にて算出したノードタイプ間ごとの推薦スコアを記載する（ただし，いずれのカテゴリにおいても一度も出現しなかった 3 ノード部分グラフパターン（12)，(13) については表記を割愛した).

時刻 t_1 (= 2007 年 1 月）までにおける QA ネットワーク Q_1 に含まれる 3 ノード部分グラフを抽出する（ステップ 4).

さらに，ノードタイプ間毎の推薦スコアと，Q_1 に含まれる 3 ノード部分グラフにどのユーザが含まれるかに関する情報を用いて，2 ユーザ間の推薦スコアを計算する（ステップ 5). ただし，ここでスコアを算出する 2 ユーザ（$v \rightarrow u$, すなわち質問者 u と回答者 v)）を以下のように限定する.

(1) u, v ともに時刻 t_0 (= 2006 年 1 月)から時刻 t_1 (= 2007 年 1 月）の期間で一度は質問ないし回答しているユーザ
(2) u, v ともに Q_1 に含まれる 3 ノード部分グラフに含まれるユーザ
(3) u はt_1 (= 2007 年 1 月）から t_1 (= 2007 年 8 月）に実際に質問したユーザ

(1) の根拠としては，t_0 以前には活動していたが t_0 以降まったく活動していないユーザが t_1 (= 2007 年1 月）以降に何らかの質問に対して回答する確率は低いと考えられるためである．(2) の根拠としては t_1 (= 2007 年1 月）以降にしか活動していないユーザは Q_1 に含まれる 3 ノード部分グラフには含まれないため，提案手法では予測できないからである．(3) の根拠としては，本提案手法においては質問者集合はあくまで入力であるという前提条件に基づく．実際に質問していない質問者の質問を回答者に推薦することはできない.

本来 v を入力として v が回答しそうな質問の質問者を求める問題だが, 今回は回答が発生しそうな 2 ユーザの組合せについて本提案手法によりスコアリングし，そのスコアの高い順でランキングした際上位何件が正解か調査することにより提案手法の精度を評価する. ただし, ここで本提案手法が v を入力としたときの u のランキングにも使えることは自明である.

5.2 正解セット

正解セットとして今回用いたのは，時刻 t_1 (= 2007 年 1 月)から t_2 (= 2007 年 8 月）までの期間で，(1)～(3) の条件を満たし，実際に回答が発生した 2 ユーザの組合せである．各カテゴリにおける (1)～(3) を満たす正解の候補となる 2 ユーザの組合せ（A) の数と，時刻 t_1 (= 2007 年 1 月）から t_2 (= 2007 年 8 月)までの期間で回答が発生した 2 ユーザの組合せ（R) 数，正解セット（$A \cap R$) の数それぞれを，表 2 に示す.

表 3　正解セットの数
Table 3　The numbers of pairs where an answer occurs

カテゴリ	候補（A)	回答発生（R)	正解（$A \cap R$)
自転車	38777	3063	245
スイーツ	1497	1308	10
海外生活	7398	710	43
クラシック	13076	1394	111
手芸	7774	1114	48
PHP	41161	1996	156
音楽	9662	1769	54
栄養	3683	1346	10
留学	11168	1778	37
Java	15023	1541	71

表に示す通り，時刻 t_1 (= 2007 年 1 月）から t_2 (= 2007 年 8 月）までの期間で回答が発生した 2 ユーザの組合せの中で (1)～(3) を満たすものはわずかである．これは，t_1 以降に新規参加したユーザが多いためである. 本提案手法では, そのようなユーザを含む 2 ユーザの組合せを評価することはできない.

5.3 評価尺度

4 節で説明した通り, 本提案手法は回答者集合 V の 1 人 1 人 v を入力とし, 適切な U の部分集合を出力とする検索のランキングに帰着する．ただし，今回の評価実験におけるタスクは本来の v の回答しそうな u の検索ではなく，2 ユーザの組合せ（u, v) $\in U \times V$ のうち実際に期間 t_1 から t_2 までに回答が発生した 2 ユーザの検索になる.

本提案手法の有効性を評価するため，本提案手法にてランキングされた上位 250，500，1000 個の 2 ユーザの組合せのうち，実際に回答が発生した組合せの数を求め，それぞれの実測値を，2 ユーザの組合せを完全にランダムに選択した際の正解率(すなわち正解数/候補数) から算出した期待値と比較する.

5.4 実験結果

表 4　正解数
Table 4 The numbers of true positive

カテゴリ	@250		@500		@1000	
	実測値	期待値	実測値	期待値	実測値	期待値
自転車	39	1.58	46	3.16	70	6.32
スイーツ	4	1.67	6	3.34	9	6.68
海外生活	12	1.45	18	2.91	22	5.81
クラシック	23	2.12	37	4.24	54	8.49
手芸	17	1.54	22	3.09	27	6.17
PHP	42	0.95	55	1.90	71	3.79
音楽(教科)	10	1.40	15	2.80	22	5.59
栄養	8	0.68	9	1.36	9	2.72
留学	10	0.83	16	1.66	22	3.31
Java	16	1.18	21	2.36	25	4.73

表 3 に実験結果を示す．今，いずれのカテゴリにおいてもランダムに選択された 2 ユーザよりも本提案手法により上位にランクされる 2 ユーザの方に正解が多いことがわかる．今，その差が統計的に有意なものであるか否かを判定す

るために，χ^2検定を行う．まず，ランダムに選択された 2 ユーザと本提案手法により上位にランクされる 2 ユーザとで正解の数に差がないと帰無仮説を立てる．次に，その帰無仮説が採択される確率を算出する(採択確率を表 4 に示す)．結果，有意水準を 1% とすると，スイーツを除くすべてのカテゴリにおいて前述の仮説が棄却されることがわかった．すなわち，本提案手法により上位にランクされた 2 ユーザ間には実際に回答の発生した 2 ユーザが多いことがわかる．

表 5　χ^2検定における帰無仮説の採択率
Table 5　Acceptance rates of null hypotheses in the χ^2 tests

カテゴリ	@250	@500	@1000
自転車	5.73×10^{-196}	3.99×10^{-129}	1.91×10^{-142}
スイーツ	7.04×10^{-2}	1.44×10^{-1}	3.68×10^{-1}
海外生活	1.54×10^{-18}	5.95×10^{-19}	1.62×10^{-11}
クラシック	5.05×10^{-47}	1.83×10^{-57}	1.89×10^{-55}
手芸	7.79×10^{-36}	2.95×10^{-27}	4.04×10^{-17}
PHP	0.00	0.00	3.69×10^{-262}
音楽(教科)	2.93×10^{-13}	2.46×10^{-13}	3.40×10^{-12}
栄養	6.17×10^{-19}	5.37×10^{-11}	1.37×10^{-4}
留学	5.52×10^{-24}	5.77×10^{-29}	7.82×10^{-25}
Java	1.80×10^{-42}	6.00×10^{-34}	9.43×10^{-21}

5.5 考察

実際に回答の発生した 2 ユーザが多いということは，少なくとも上位における 2 ユーザの組合せ (u, v) において，「回答者 v にとって質問者 u の質問はより回答するのに適切な質問である」と推測することができる．したがって，本提案手法により上位にランクされる 2 ユーザ (u, v) に対して回答者 v に質問者 u の質問を推薦すれば，同じ回答者 v にランダムに選択した質問を推薦するよりは回答をつけてくれる可能性が高いとも考えられる．

6. まとめと今後の課題

本研究では，ネットワークの成長に関する制約をその部分グラフのパターンに着目することにより推測し，その制約からリンク予測を行うという手法を提案した．我々は特に 3 ノード部分グラフに着目し，その対称性からノードを 30 種類のタイプに分類，さらにそれらの間で回答が発生する可能性のある箇所を 54 種類に分けた．54 種類のノードタイプ間のそれぞれの回答の発生しやすさに関して評価し，それを特定の 2 ユーザ間の回答の発生しやすさに集約する手法も提案した．

本稿では提案手法の評価を行うために，QA サイトを対象として，2007 年 1 月から同年 8 月の間に回答が発生した 2 ユーザを検索するタスクについて，本提案手法と完全にランダム選択した場合とで比較評価実験を行った．結果，本提案手法が有効であることがわかった．このことは本提案手法が回答者 v に適切な質問者 u の質問を推薦するのに有効である可能性を示唆している．

回答者への質問推薦という観点から，今後以下のような課題に取り組んでいく予定である．

- 評価実験におけるベースラインを見直す必要がある．完全にランダムに 2 ユーザを選択するのではなく，既存のリンク予測アルゴリズムを用いる．
- 回答者への質問推薦のプロトタイプを作成，ユーザ実験を行う．
- 回答者の知識に基づく手法を考案する．

[文献]

[1] R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, and U. Alon. Network Motifs: Simple Building Blocks of Complex Networks. Science, Vol. 298, No.5594, pp. 824–827, 2002.
[2] S. Wernicke. Efficient Detection of Network Motifs. IEEE/ACM Transactions on Computational Biology and Bioinfomatics, pp. 347–359, 2006.
[3] 高田寛喜, 山田武士, 上田修功. ノードの機能特性に基づくクラスタリング. ネットワーク生態学シンポジウム2008, pp. 120-124, 2008.
[4] MEJ Newman. Clustering and preferential attachment in growing networks. Physical Review E, Vol. 64, No. 2, p. 25102, 2001.
[5] R. Baeza-Yates, B. Ribeiro-Neto, et al. Modern information retrieval. Addison-Wesley Harlow, England, 1999.
[6] D. Liben-Nowell and J. Kleinberg. The Link-Prediction Problem for Social Networks. Journal-American Society For Information Science and Technology, Vol. 58, No. 7, p. 1019, 2007.
[7] L.A. Adamic and E. Adar. Friends and neighbors on the Web. Social Networks, Vol. 25, No. 3, pp. 211–230, 2003.
[8] L. Katz. A new status index derived from sociometric analysis. Psychometrika, Vol. 18, No. 1, pp. 39–43, 1953.
[9] H. Welser, E. Gleave, D. Fisher, and M. Smith. Visualizing the Signatures of Social Roles in Online Discussion Groups. The Journal of Social Structure, Vol. 8, No. 2, 2007.
[10] 甲谷優, 川島晴美, 藤村考. QA サイトにおける質問応答グラフの成長パターン分析. 情報処理学会研究報告. データベース・システム研究会報告, Vol. 2008, No. 88, pp. 247-252, 2008.
[11] L.A. Adamic, J. Zhang, E. Bakshy, and M.S. Ackerman. Knowledge sharing and Yahoo Answers: Everyone knows something. Proceedings of the 17th international conference on World Wide Web, 2008.
[12] 大平雅雄, まつ本真佑, 伊原彰紀, 松本健一. オープンメディアを活用した知識コミュニティのデザインに関する考察. 知識共有コミュニティワークショップ, pp. 1-10, 2008.
[13] 村田剛志, 森保さき子, 池谷智行. 社会ネットワークとしてのYahoo!知恵袋. 知識共有コミュニティワークショップ, pp. 11-18, 2008.

甲谷 優　Yutaka KABUTOYA
NTT サイバーソリューション研究所 所属．2008 年京都大学大学院情報学研究科博士前期課程修了．同年，日本電信電話株式会社入社．Web マイニングの研究開発に従事．日本データベース学会，人工知能学会，言語処理学会，各会員．

川島 晴美　Harumi KAWASHIMA
NTT サイバーソリューション研究所 主任研究員．1990 年山梨大学大学院工学研究科博士前期課程修了．同年，日本電信電話株式会社入社．現在インターネットからの話題情報抽出技術の研究開発に従事．電子情報通信学会会員．

藤村 考　Ko FUJIMURA
NTT サイバーソリューション研究所 主幹研究員．電気通信大学大学院情報システム学研究科客員教授．1989 年 北海道大学大学院工学研究科博士課程修了．同年，日本電信電話株式会社入社．トランザクション処理記述言語，汎用電子チケットシステム，電子決済システム，blog マイニングの研究開発に従事．博士(工学)．情報処理学会，電子情報通信学会，日本社会情報学会，各会員．