

コミュニティQAにおける質問・回答間の依存関係の評価

香川 雄一[†] 島田 諭[†] 神門 典子^{††} 佐藤 哲司[†]

[†] 筑波大学大学院図書館情報メディア研究科 〒305-8550 茨城県つくば市春日 1-2

^{††} 国立情報学研究所 〒101-8430 東京都千代田区一ツ橋 2-1-2

E-mail: [†]{kagawa,sat,satoh}@ce.slis.tsukuba.ac.jp, ^{††}kando@nii.ac.jp

あらまし コミュニティQAでは、1つの質問に複数の回答が投稿されるのが一般的である。このため、2個目以降の回答を投稿する回答者は、質問だけでなく既に投稿された回答を見てから自分の回答を投稿すると考えられる。本研究では、質問に連なる回答間の依存関係をネットワーク構造で表現し、質問・回答間の依存関係を定量的に評価するという新しいアプローチを試みる。質問・回答間の依存関係を表すネットワーク図を作成するための規則を作成し、ネットワーク図を作成する実験を行なった。2名の実験者によって作成されたネットワーク図を分析し、実験者に依らず安定して質問・回答間の依存関係を抽出できるかの評価を行ったので報告する。

キーワード コミュニティQA, 談話構造解析

Dependence relationships among Question and Answers in Community QA

Yuichi KAGAWA[†], Satoshi SHIMADA[†], Noriko KANDO^{††}, and Tetsuji SATOH[†]

[†] Graduate School of Library and Information Science and Media Studies, University of Tsukuba
1-2, Kasuga, Tsukuba, Ibaraki, 305-8550 Japan

^{††} National Institute of Informatics

2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo, 101-8430, Japan

E-mail: [†]{kagawa,sat,satoh}@ce.slis.tsukuba.ac.jp, ^{††}kando@nii.ac.jp

Abstract In Community QA, several answers are usually posted in each question. Therefore, it is supposed that answerers post their own answers after seeing not only the question but also answers posted before. In this paper we express dependence relationships among question and answers with network structure and try to evaluate the dependence relationships. We make the rule of making network charts indicating dependence relationships and experiment making network charts. We analyze network charts made by two people. We report on the result of evaluation about whether extracting dependence relationships among question and answers can be done or not independent of who makes.

Key words Community QA, Discourse Structure Analysis

1. はじめに

近年、電子掲示板、コミュニティQAなどウェブ上の情報交換・知識共有サイトの利用が盛んである。コミュニティQAとは、電子掲示板上で利用者同士が知識や知恵を教え合うナレッジコミュニティ、知識検索サービスである[1]。例として、日本語の質問回答サイトの中で最大級のYahoo! 知恵袋^(注1)や教えて!goo^(注2)などが挙げられる。

コミュニティQAでは、質問に対する直接的な回答だけでなく、既に投稿された回答への補足や反論など追記型の投稿がなされることが多い。そのため、1つの質問に複数の回答が投稿されるのが一般的である。2つ目以降の回答を投稿する回答者は、質問だけでなく自分より先に投稿された回答を見て、回答するかどうかの判断も含めて自分の回答の内容を決めると考えられる。例えば、遠距離恋愛を続けるべきかどうかの質問があった時に、続けるべきと言う先行した回答に対して、自らの体験を補足する、あるいは、続けるべきとした回答の根拠が曖昧として、慎重な行動を取るべきとする回答が投稿されるなど、先の回答を参照しながら賛同や反論を示す回答などがある。

(注1): <http://chiebukuro.yahoo.co.jp/>

(注2): <http://oshiete.goo.ne.jp/>

従来、質問のタイプ分類やベストアンサーの推定などコミュニティQAに関する研究は多いが、質問・回答間の依存関係に着目した研究はあまり知られていない。そこで本研究では、質問に連なる回答間の依存関係をネットワーク構造で表現し、質問・回答間の依存関係を定量的に評価するという新しいアプローチを試みる。これにより、質問・回答間の依存関係と各回答が持つ役割を明らかにすることを目指す。また、依存関係をネットワーク図で表現することで、質問タイプの類型化や、ベストアンサーとなるための要件と良質な回答の特徴を推定することに利用することを目指す。ベストアンサーとなるための要件や良質な回答の特徴としては、既存の回答をまとめるような回答や、新しい観点を提示している回答などが考えられる。

本論文の構成は以下の通りである。2章で関連研究について述べる。3章で質問・回答間の依存関係について述べる。4章で実験概要について述べたあと、5章で考察を示し、6章でまとめを述べる。

2. 関連研究

コミュニティQAにおいて、質問や回答のつながりについて、様々な視点から着目したアプローチがされている。島田ら[2]は、質問とその質問に対する一連の回答をチャンクと称し、同一のチャンクに出現する回答者の共起関係から興味分野の重なりを抽出し、投稿記事を関連付ける手法を提案している。甲谷ら[3]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するQAネットワークに着目している。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。佐藤ら[4]は、知識共有サイトでは投稿活動の活性化が、参加者間の知識共有の促進へと繋がると考え、参加者の投稿行動をネットワーク構造に表現し、コミュニティに対する参加者の貢献度を推定する手法を提案している。参加者の投稿行動として「質問に対する第一回答（FA）」と「ベスト回答（BA）」に着目して、ネットワークを構成し、この2つのネットワークに対して提案手法を適用している。鈴木ら[5]は、言語表現から自己中心性を推定する手法を用いて、質問回答サイトの対話における意思疎通のズレを推定する手法を提案している。

これらのように、質問記事同士の関連付けや、質問者や回答者に着目したネットワークの研究は多くなされているが、1つの質問とそれに連なる回答（本研究では、チャンクと定義する）の依存関係についての研究は知られていない。

3. 質問・回答間の依存関係

本章では、複数の質問記事に対して、各チャンクにおける質問・回答間の依存関係をネットワーク図で表現するための具体的な方法について説明する。

テキスト中の談話単位間の関係を解析し、テキストの構造を明らかにすることは、一般的に談話構造解析と言われている[6]。横山ら[6]が述べているように、談話構造解析を、機械学習を用いて行う方法があるが、コンピューターが自動で解析を行うことは非常に困難であり、解析精度の点でまだ実用レベルに達

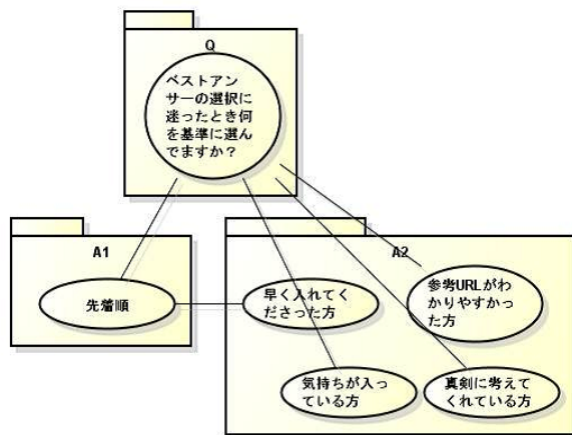


図1 質問・回答間のネットワーク図の例

していない。本研究では、質問・回答間の依存関係を示すための手段としてネットワーク図を用い、ネットワーク図作成のための規則を手動で作り出す。

ネットワーク図の作成規則（初訂版）

- (1) 質問文を読んで、「質問の聞きたいこと、知りたいこと」を質問の観点として抜き出し、質問内のノードとする。
- (2) 回答文を読んで、「各回答ごとに応えていること、応えようとしていること」を一つ一つ回答の観点として分解し、各回答内のノードする。
- (3) 作成したノードにおいて、質問内のノードと対応している回答の観点のノードの間で無向のリンクを張る。
- (4) また、読み取った回答の観点のノード間で対応しているもの同士で無向のリンクを張る。

以上の作業を行うことで、質問・回答間の関係を再構成し、ネットワーク図を作成する。例えば、「ベストアンサーの選択に迷ったとき何を基準に選んでいますか？」という質問に対して、時系列順に「迷ったら先着順です。」「早く入れてくださった方や参考URLがわかりやすかった方、気持ちが入っていて、真剣に考えてくれている方、など、ですかねえ。」という回答があったとする。この場合、質問の聞きたいこと、知りたいこととして、「ベストアンサーの選択に迷ったとき何を基準に選んでいますか？」というノードが抽出される。質問に対する回答として前者の回答は「先着順」、後者の回答は「早く入れてくださった方」「参考URLがわかりやすかった方」「気持ちが入っている方」「真剣に考えてくれている方」というノードがそれぞれ抽出できる。これらの抽出したノードに基づいて同じ観点をリンクでつなぐとき、回答の「先着順」「参考URLがわかりやすかった方」「気持ちが入っていて、真剣に考えてくれている方」というノードはそれぞれ質問「ベストアンサーの選択に迷ったとき何を基準に選んでいますか？」に対して最初に出た観点なので、直接質問と回答のノード間でリンクを張る。ここで、回答の「早く入れてくださった方」というノードは、「回答の早さ」という観点で「先着順」というノードと同じであると考えられるので、両ノード間にリンクを張る。同様に、後から

表 1 NTCIR-8 CQA において投稿数の多いカテゴリと投稿数

カテゴリ名	投稿数
恋愛相談, 人間関係の悩み,	704
Yahoo!オークション	636
パソコン, 周辺機器	211
政治, 社会問題	184

投稿された回答に, 既に投稿された回答と同じ観点がある場合は, 直近の回答のノードとリンクを張る. 以上の結果を, 図 1 に示す.

4. 予備実験概要

次に, 実験者に依らず安定して質問・回答間の依存関係を抽出できるかを評価するために, 前章で述べたネットワーク図の作成規則に基づいて, 複数人によって手動でネットワーク図を作成する予備実験を行う.

4.1 分析対象とするデータ

本研究では, 大学共同利用機関法人国立情報研究所が提供をしている, Yahoo!知恵袋のデータを用いる. 実験では, NTCIR-8 CQA (コミュニティQA テストコレクション)^(注3)に含まれる質問と回答を用いる. これは 2004 年 4 月から 2005 年 10 月に投稿された質問から無作為に 1500 件を抽出したもので, 全 14 のカテゴリからなる. 本評価では, 表 1 に示すように, 特に投稿数の多い「恋愛相談, 人間関係の悩み」, 「Yahoo!オークション」, 「パソコン, 周辺機器」, 「政治, 社会問題」の 4 カテゴリから 10 件ずつランダムに抽出する. さらにカテゴリによる依存を調査するために, 前者で 10 件ずつ取り出したものとは別に, 全 14 のカテゴリから, 各 2 件ずつ抽出し, 合計 68 件を対象とする.

4.2 実験の手順

実験者に, チャンクごとに質問と回答を時系列順に提示する. 実験者はこれを見ながら, ネットワーク図を作成していく. 前章で述べたように実験者は, 読み取ったことをノードとして書き込み, ノード間で観点が対応しているもの同士をリンクで結ぶ. 以上の手順に基づいて, 実験者を 2 名として実験を行った. 実験者の理解を高めるためのプロセスとして, 作業→インタビュー→作業…という流れで実験を行なった. インタビューでは, 実験者が実験中に特に困った問題や作成規則に対する疑問を調査した.

4.3 ネットワーク図の一致度

本稿では, 実験者に依らず安定して質問・回答間の依存関係を抽出できるかの評価を行うことを目的に, 2 名が作成したネットワーク図の一致度を調査する.

ノード一致度が高かったネットワーク図の例を図 2, ノード一致度が低かったネットワーク図の例を図 3 に示す. 図 2, 図 3 に示すように, 2 名が作成したネットワーク図において, 同一の質問内のノード, 同一の各回答内のノードで, それぞれ同じ観点であると考えられるノードをマークを変えて示す. これ

らをそれぞれ質問の一致したノード, 回答の一致したノードとする. また, 2 名が作成したネットワーク図において, 同じ観点としたノード間で 2 名とも張ったリンクを一致したリンクとする.

分析対象としたデータ 68 件のうち, 2 名ともネットワーク図を作成することができた 52 件のデータにおける, ノード, リンクの一致度を Jaccard 係数を用いて求めた. 残りの 16 件については, 1 人のみが作成あるいは 2 人とも作成しなかった. Jaccard 係数は次の式を用いて求められる. X, Y は実験者 A, 実験者 B が抽出したノード, リンク数を示す.

$$Jaccard = \frac{X \cap Y}{X \cup Y}$$

質問のノードの一致度, 回答のノードの一致度, 全てのリンクを対象としたときのリンクの一致度, 一致したノード内のリンクのみを対象としたときのリンクの一致度の平均値を表 2 に示す. また, 回答のノードの一致度と全てのリンクを対象としたときのリンクの一致度の相関を図 4 に, 回答のノードの一致度と一致したノード内のリンクのみを対象としたときのリンクの一致度の相関を図 5 に示す.

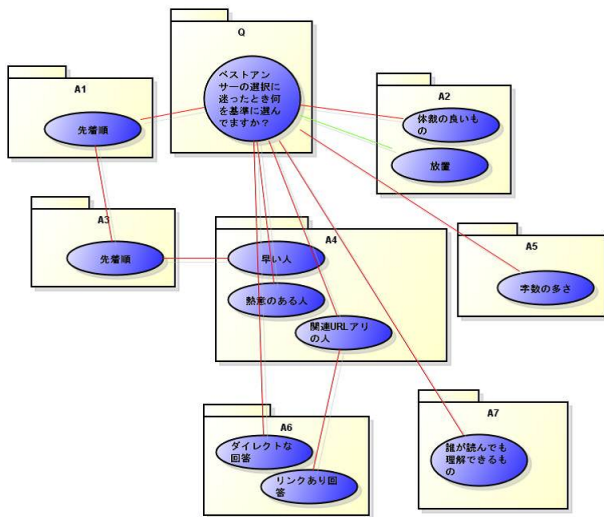
表 2 ノード, リンクの一致度の平均値

ノード, リンク	一致度の平均値
質問のノード	0.941
回答のノード	0.587
全てのリンクを対象としたときのリンク	0.371
一致したノード内のリンクを対象としたときのリンク	0.628

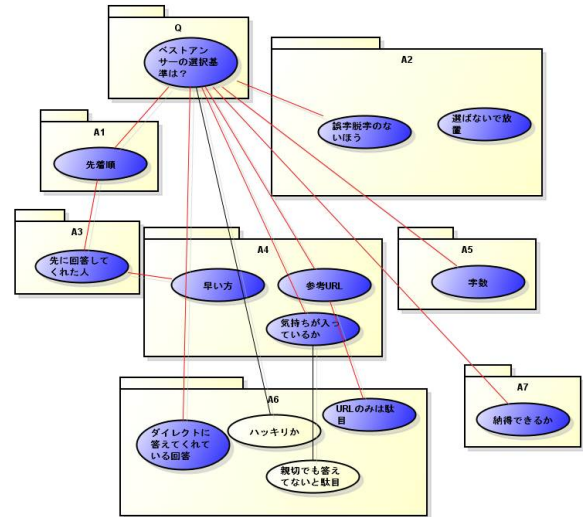
4.4 考察

図 2 のように, ノード一致度が高かった例は「先着順」「体裁の良いもの」など文ではなく単語で回答することができる質問に多いと考えられる. また, 図 3 のように, ノード一致度が低かった例は, 単語では回答しにくく, 文で回答することが多い質問に多いと考えられる. 表 2 より, 質問のノードに関しては, 一致度が大きく, 実験者に依らず安定して質問の聞きたいこと, 知りたいことを抜き出すことができることが分かった. 回答のノードに関しては, 表 2 より, 一致度の平均値は 0.587 で, 質問のノードより実験者によって, 各回答ごとに応えていること, 応えようとしていることの読み取り方のちがいが大きいことが分かった. リンクの一致度に関しては, 表 2 より, 平均で全てのリンクを対象とすると, 0.371 であるが一致したノード内のリンクを対象にすると 0.628 にまで向上している. 一致していない回答のノードが多いことを考えると, 回答のノードの抽出の方法を変えて実験を行えば, リンクの一致度が上がり, ネットワーク図の一致度はかなり上がると考えられる. ノードの抽出の方法として, 実験者に自由に作文などさせてしまうと, 一致度を図る際などに曖昧さが残るので, 記事中に出てくる用語をそのまま抽出し, ノードとするという方法が考えられる.

(注3): http://research.nii.ac.jp/ntcir/permission/ntcir-8/perm-ja-CQA.html#cqa_task

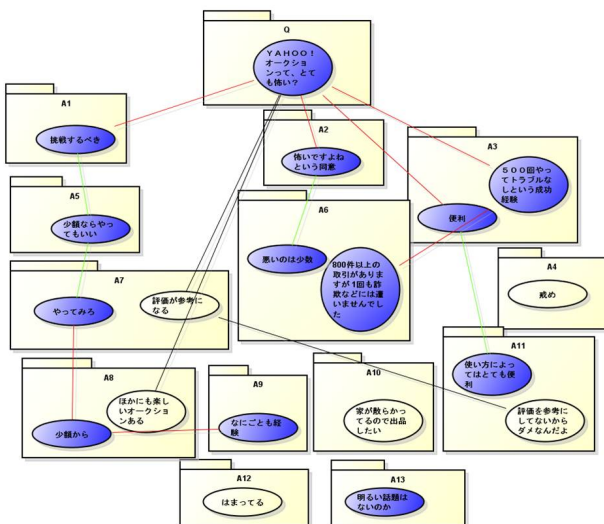


(a) 実験者 A が作成

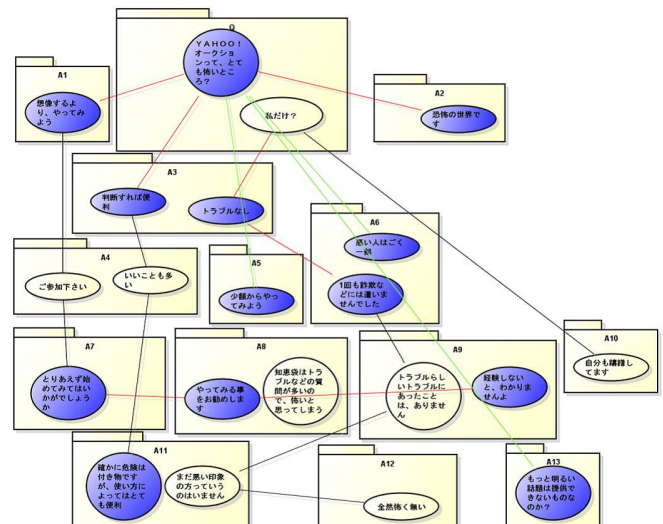


(b) 実験者 B が作成

図 2 ノード一貫性が高かったネットワーク図の例



(a) 実験者 A が作成



(b) 実験者 B が作成

図 3 ノード一貫性が低かったネットワーク図の例

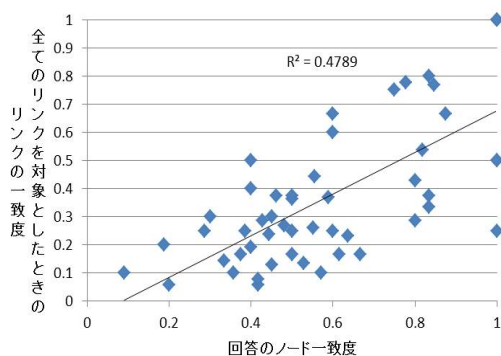


図 4 相関が見られた回答のノード一貫性とリンクの一貫性

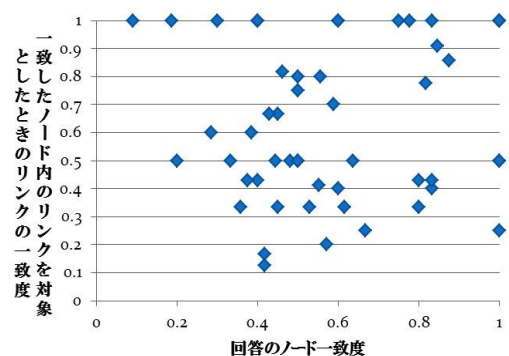


図 5 相関が見られなかった回答のノード一貫性とリンクの一貫性

5. 本実験概要

次に、前章の予備実験をから得られた知見を元に得られた作

成規則を用いて、再び複数人によって手動でネットワーク図を作成する実験を行なった。

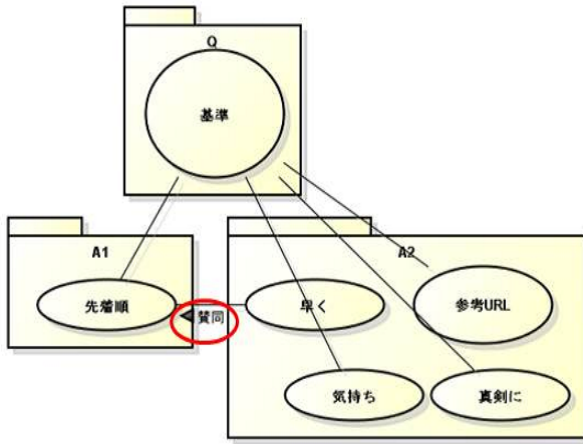


図 6 質問・回答間のネットワーク図の例

5.1 分析対象となるデータ

本実験では、NTCIR-8 CQA（コミュニティQA テストコレクション）に含まれる質問をのうち、特に回答数が多く、回答間の依存関係が多いと思われる「恋愛相談、人間関係の悩み」、 「Yahoo!オークション」の2カテゴリから QID 順に 108 件抽出する。実験者に依らず、ネットワーク図を作成できるようにすることを目的とするため、実験者は、予備実験とは異なる 2 名で実験を行った。

5.2 実験の手順

本実験で実験者に提示した改訂後の作成規則を示す。

ネットワーク図の作成規則（改訂版）

ノードの抽出（観点の切り出し）

(1) 質問文を読んで、「質問の聞きたいこと、知りたいこと」を質問の観点として抜き出し、質問内のノードとする。

(2) 回答文を読んで、「各回答ごとに応えていること、応えようとしていること」を一つ一つ回答の観点として分解し、各回答内のノードとする。

・記事中に出てくる用語を抽出しコピー、ペーストでノードとする

・用語は記事中の出現位置が特定できる範囲で最短の文字列とする

・動詞、形容詞などは語幹を抽出（厳密でなくても良い）

・複合名詞などは、2 語程度の連接まで許容する

・離れた 2 カ所の用語を取る場合は、スペースで並べる

・観点として切り出すべきか悩んだ場合は、切り出してノードにする。（迷ったらノードとし、できるだけ多くのノードを抽出する）

・質問と直接関係ないノードでも他の回答とつながりがあれば、ノードを作成してリンクを張る

リンク付与

(1) 作成したノードにおいて、質問内のノードと対応して関連のある回答の観点のノードの間で無向のリンクを張る。

(2) また、読み取った回答の観点のノード間で対応しているもの同士でリンクを張る。

・出リンク数、入リンク数とも 1 に限定はしない（1 つのノードから複数のリンクが出て入ってもよい）

・ある観点（ノード A）と同じ観点の回答であると考えられる場合は、後から書かれた回答記事中の観点（ノード B）からノード A にリンクを張る

・リンクを張るノードの文字列は一致している必要はない

・リンク付与後、リンクに名前をつける。以下の 4 つから選択する（当てはまらない場合は無記入）

賛同：既存の観点について賛成、同意することを述べる

反論：既存の観点について違ったこと、対立すること矛盾することを述べる

具体化（例示）：既存の観点を、具体的な例を挙げて説明するなどして深める・強化する

抽象化（まとめ）：既存の観点を要約したり、集約する

本実験では、新たに「リンクに名前をつける（ラベル付）」という手順を追加する。これを行うことで、リンクを付与したノード間の関係を示す。例えば、図 1 のように「先着順」というノードと「早く入れてくださった方」というノードは、「回答の早さ」という観点で同じで、さらに「回答の早い方」という点で同意なのでリンクのラベルとして「賛同」をつける。図 1 の例に対して、以上の手順を追加した結果を図 6 に示す。

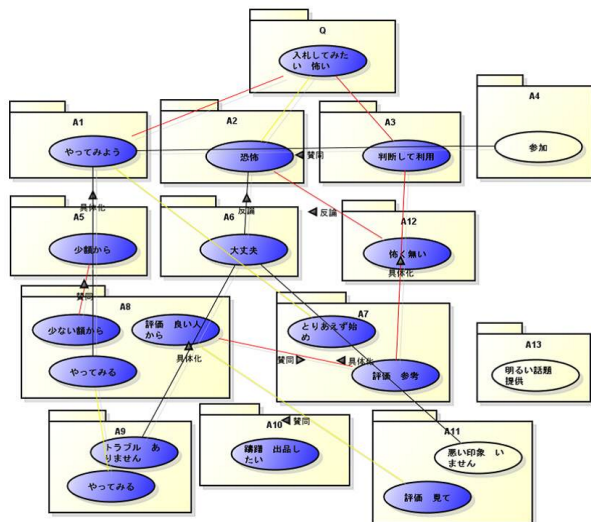
5.3 ネットワーク図の一致度

予備実験と同様に、実験者に依らず安定して質問・回答間の依存関係を抽出できるかの評価を行うことを目的に、2 名が作成したネットワーク図の一致度を調査する。ノードの一致度については、共通の文字列を含めばノードが一致したとみなし、共通の文字列を含まなければノードは一致していないとみなす。図 6 に、図 3 と同じ質問記事で予備実験時から一致度が高くなった例を示す。

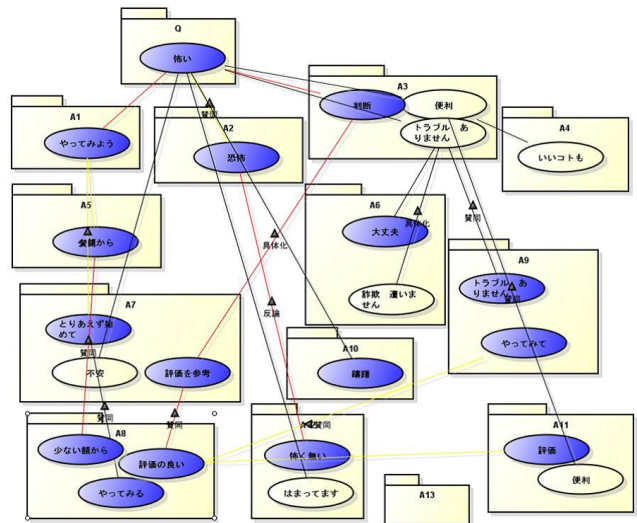
分析対象としたデータ 108 件のうち、2 名ともネットワーク図を作成することができた 81 件のデータにおける、ノード、リンクの一致度を Jaccard 係数を用いて求め、その結果を表 3 に示す。残りの 27 件については、1 人のみが作成あるいは 2 人とも作成しなかった。

表 3 ノード、リンクの一致度の平均値

ノード、リンク	一致度の平均値
質問のノード	0.597
回答のノード	0.551
全てのリンクを対象としたときのリンク	0.292
一致したノード内のリンクを対象としたときのリンク	0.627



(a) 実験者 A が作成



(b) 実験者 B が作成

図 7 ノード一致度が高くなったネットワーク図の例

6. 考 察

表 3 より、質問のノードに関しては、予備実験時と比較して一致度が大きく下がっている。これは、例えば「ベストアンサーの選択に迷ったとき何を基準に選んでいますか？」という質問があったときに、「基準」「何を 選んでいますか」というように、実験者に依って、質問の内容の部分抽出する人と直接問いかけの部分抽出する人に分かれてしまうからだと考えられる。回答のノードに関しては、一致度の平均値は 0.551 で、予備実験時と比較して大きな差異はなかった。予備実験と本実験では利用した質問記事が異なるので直接は比較できないが、本実験では回答数の多い質問記事を中心に利用したので、予備実験時と比較して精度は良くなったと考えられる。図 6 のように、観点を単語単位でできるだけ多くの抽出させることで一致度が高くなったと考えられる。リンクの一致度に関しては、表 3 より、全てのリンクを対象とすると、0.292 であるが一致したノード内のリンクを対象にすると 0.627 にまで向上しており、予備実験時とほぼ同じ値であるので、全てのリンクを対象としたときのリンクの一致度が低いのは質問のノードの一致度が低いことが原因であったと考えられる。今後の課題として、ネットワーク図の作成規則を改良することが挙げられる。具体的には、質問のノードを抽出する際には、質問の内容の部分と直接問いかけの部分抽出する。さらに、回答のノードを抽出する際も同様に、回答の大まかな内容の部分と直接回答している部分を抽出する。

7. おわりに

本研究では、質問に連なる回答間の依存関係をネットワーク構造で表現し、質問・回答間の依存関係を定量的に評価するという新しいアプローチを試みた。まず、質問・回答間の依存関係を表すネットワーク図作成のための規則を手動で作り出した。次に、実験者に依らず安定して質問・回答間の依存関係を抽出

できるかを評価するために、ネットワーク図の作成規則に基づいて、2 名の実験者によって手動でネットワーク図を作成する実験を行った。その結果、質問のノードに関しては、一致度が大きく、実験者に依らず安定して質問の聞きたいこと、知りたいことを抜き出すことができることが分かった。回答のノードに関しては、一致度は質問のノードよりは低く、質問のノードより実験者によって、各回答ごとに応えていること、応えようとしていることの読み取り方のちがいが大きいことが分かった。リンクの一致度に関しては、全てのリンクを対象とすると低いが、一致したノード内のリンクを対象にすると大きく向上することが明らかとなった。

今後の課題として、ネットワーク図の作成規則を改良することが挙げられる。また、ネットワーク図を用いて、質問タイプの類型化や、ベストアンサーとなるための要件と良質な回答の特徴の推定など様々な側面から評価を行うことが挙げられる。

謝 辞

本研究の一部は科研費 (21500091) の助成を受けたものである。本研究の実装・評価に際し、大学共同利用機関法人国立情報学研究所から提供を受けた Yahoo!知恵袋のデータを利用している。ここに記して謝意を示す。

文 献

- [1] 三浦麻子, 川浦康至, 地福節子, 大瀧直子, 岡本真. 知識共有コミュニティを創り出す人たち. 人工知能学会全国大会大会論文集, pp. 1-4, Jun.2006.
- [2] 島田諭, 福原知宏, 佐藤哲司. 質問回答サイトにおける利用者行動に基づく記事の関連付け手法の検討. 電子情報通信学会 第 19 回データ工学ワークショップ (DEWS2008), No. B6-5, 2008.
- [3] 甲谷優, 川島晴美, 藤村考. 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. 日本データベース学会論文誌 8(1), pp. 89-94, Jun.2009.
- [4] 佐藤弘樹, 島田諭, 伏見卓恭, 福原知宏, 斉藤和巳, 佐藤哲司. 知

識共有サイトにおける参加者の貢献度に着目したコミュニティ分析手法. 第2回データ工学と情報マネジメントに関するフォーラム (DEIM2010), No. B1-1, 2010.

- [5] 鈴木信雄, 津田和彦. 自己中心性に基づく質問回答サイトにおける意思疎通のズレの表現. 情報処理学会全国大会講演論文集 第71回, pp. "2-71"-"2-72", 2009.
- [6] 横山憲司, 難波英嗣, 奥村学. Support vector machine を用いた談話構造解析. 自然言語処理研究会報告 2003(23), pp. 193-200, 2003.