共起制限と係り受け構造グラフを利用した宿泊レビューの特徴表現抽出手法

Koji Tanaka
Hitachi Government & Public Sector Systems, Ltd
Tokyo, Japan
koji.tanaka.st@hitachi.com

Koichi Tsujii
Nippon Travel Agency Corpolation LTD.
Tokyo, Japan
tsujii@gmail.com

Takashi Ikoma
University of Tsukuba
Tokyo, Japan
ikoma@gssm.otsuka.tsukuba.ac.jp

Akiyuki Sekiguchi
Renesas Electronics Corporation.
Tokyo, Japan
akiyuki.sekiguchi.pd@renesas.com

Kazuhiko Tsuda
University of Tsukuba

Tokyo, Japan
tsuda@gssm.otsuka.tsukuba.ac.jp

Abstract—インターネット宿泊予約サイトの宿泊レビューは宿泊施設選択の参考情報として重要視され、マーケティングへの活用も期待されている。ホテルレビューのコメントは記載のバリエーションが多い。このため、類義語辞書の作成が必要であった。しかしながら、類義語辞書の作成には手間がかかるといった問題や随時見直しが必要になるといった問題がある。また、予め評価対象を決めた上で、その対象に関する類義語を設定しなければならないため、想定していない評価対象に関しての分析が困難になるという問題がある。

このため、本研究では、まず、共起制限と係り受け構造を利用して、感想コメントをグラフする。そして、そのグラフをクラスタリングすることで、特徴表現を抽出した。これにより、予め評価対象を設定することや類義語辞書を作成することなく、宿泊レビュー中の感想コメントから評価に関する特徴表現を抽出することができた。

Keywords—Japanese text analysis, Thesaurus, graph, Hotel Reviews

I. INTRODUCTION

インターネット宿泊予約サイトの登場で，宿泊検討者は 24 時間いつでも宿泊施設の予約ができるようになり，利用者数はこの 10 年で 10 倍を超える成長を示している[1]．宿泊検討者は最も良い条件の宿泊施設を選ぶために宿泊予約サイトが提供する様々な情報を活用し情報を収集する．昨今の宿泊検討者は，宿泊施設の提供する広告や公式情報だけを一方的に鵜呑みにせず，サービス提供を受けた宿泊者が残す宿泊レビューを閲覧し，その評価を宿泊施設選択の参考情報として重要視している[2]．また、宿泊施設にとっても、宿泊レビューから改善点やアピールポイントを把握するなど、マーケティングに宿泊レビューを活用することが期待されている[3,4]

宿泊レビューは宿泊者のプロフィール，数値評価，感想コメントの 3 つ要素から構成され，感想コメントには宿泊者の実際の体験に基づく生の声が投稿されている．宿泊レビューは有益な情報であるが，その情報提供方法は宿泊者の評価情報を並べて掲載しているだけであるため，人気宿泊施設に膨大に集まる宿泊レビューは，通読することが困難である．また，数値評価情報は個人の主観による評価となるが，一般的に高い評点をつける宿泊者が多い．またすべての数値評価を細かく評価せず，オール 5 といった様に全ての数値評価対象を同一にする宿泊者も多く存在する．そのため，数値評価を集計するだけでは，宿泊施設の具体的な特徴を知ることは難しい．

以上のことから、宿泊レビュー中の感想コメントから評価に関する特徴表現を抽出し宿泊施設選択やマーケティングに活用することが期待される。しかしながら、感想コメントは宿泊者が自由に記述する情報のた

め，表現の揺れが大きい．そのため評価対象に合わせて類義語辞書を作成し，表現を揃える処理を行う必要がある[4,5]．この類義語辞書は、サンプリングした感想コメントから目視で特徴語を抽出したり[4]、単語頻度分析による特徴語抽出を行なった結果から作成されている[6]。このため、類義語辞書の作成には手間がかかる他、宿泊レビュー投稿数が増加した場合に見直しが必要になるといった問題や、想定していない評価対象に関しては類義語辞書が作成されず、想定外の評価対象に関しての評価が困難になると言った問題がある。

本稿では、予め評価対象を設定することや類義語辞書を作成することなく、宿泊レビュー中の感想コメントから評価に関する特徴表現を抽出する手法の検討を行った結果を示す。

II. RELATED WORK

A. 宿泊レビュー分析における評価属性辞書

辻井らの研究[6]では、宿泊レビュー分析にあたり、小林らの評判分析に関する研究[7]を，宿泊レビューに適用すると共に拡張して「評価属性辞書」を作成している。

小林らの研究[7]では、評判分析に用いる肯定・否定表現の評価は，〈対象〉，評価対象の具体的な項目である〈属性〉，および〈評価〉の３要素から構成されるとしている。辻井らの研究[6]では、〈対象〉を”風呂”や“部屋”といった宿泊レビューにおける数値評価の項目とし，〈属性〉を“大浴場”や“ベッド”など具体的な評価物，〈評価〉を”広い”などの評価物の状態としている．そして，〈属性〉と〈評価〉の係り受けに対して，肯定極性・否定極性を付与した評価属性辞書を作成している。

これは、<対象>と<属性>が語の意味的な類似性を表現し、係り受け関係を用いて〈属性〉と〈評価〉の関係性を表現しているものであるといえる。

B. シソーラス

語の意味的な類似性を表現する辞書として，シソーラス辞書は，自然言語処理だけでなく幅広い研究領域で利用されてきた[8]。代表的なものとしてWordNet[9]がある。しかし，このようなシソーラス辞書の作成においては，概念を追加・更新するためには人間の手作業による膨大な手間がかかるため，最新の概念や一般的でない語彙などへの対応が難しいのが現状である.そのため，精度の高いシソーラス辞書を低コストで(半)自動的に作成する手法が必要とされている。シソーラス辞書の精度は，解析対象とするコーパスとその解析方法に強く依存するため，解析対象(コーパス)と解析アプローチともに多種多様な手法が提案されてきた。たとえば，語のクラスタリング手法を用いるもの[10]や、Web ページ間のリンク構造を利用して作成するものがある[11,12].これらの手法は、似た内容の文書間で利用される単語は、意味的な類似性をもった単語が利用されやすいことを利用している。

また、語の共起関係に基づいてシソーラス辞書を作成する研究[8]もある.語の共起関係においては、例えば、形容詞“広い”は、”部屋””風呂”など空間・場所を示すような名詞を修飾するというように、語の係り受け関係には意味的な制限があることがしられている。これを共起制限という。共起制限を利用して類義語を抽出する研究として、名詞に対して他の単語との係り受け関係の頻度を利用するもの[13]、複合語内の名詞に注目して名詞間の連接関係を利用するもの[14]がある。全ての品詞に利用して同義表現を抽出しようとする研究もある[15]

III. 共起制限と係り受け構造グラフ

Ⅱ.A で述べたように，宿泊レビュー分析では係り受け関係用いて〈属性〉と〈評価〉の関係性を表現している。また、Ⅱ.B で述べたように語の係り受け関係には共起制限があり、共起制限を利用することで語の意味的な類似性が抽出できることが期待できる。このため、本研究では、係り受け構造をグラフ構造で表現し、そのグラフをクラスタリングことで、宿泊レビュー中の感想コメントから評価に関する特徴表現を抽出する。

具体的な手順を以下及び Fig.1 に示す。

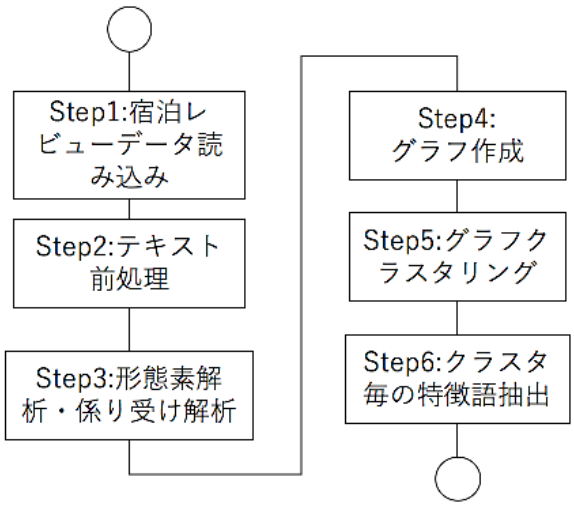

Fig.1 処理フロー

Step.1 で、宿泊レビューデータを読み込む。読み込み対象は宿泊レビューデータのうち、レビュー対象の宿泊施設名称と日本語で記載されている感想コメントである。

Step.2 では、テキストデータ(感想コメント)に対して前処理を行う。日本語では、同じ意味の文字でも全角文字と半角文字が混在するため、次に行う形態素解析で正しい結果が得られるように、これらの表記を統一する必要がある。また、長音記号「ー」とマイナス記号「−」のように、字形が似ているために誤用されることがしばしば見られ、誤用されている場合に形態素解析で正しい結果が得られない。このための修正も前処理で行う。

Step.3 では形態素解析と係り受け解析を行う。形態素解析器及び形態素辞書には MeCab[16]と mecab-ipadic-NEologd[17]を用いた。係り受け解析器には Cabocha[18]を用いた。

Step4.では、Step.3 の解析結果からグラフデータを作成する。具体的には、Fig2,3 に示すようなグラフを作成する。

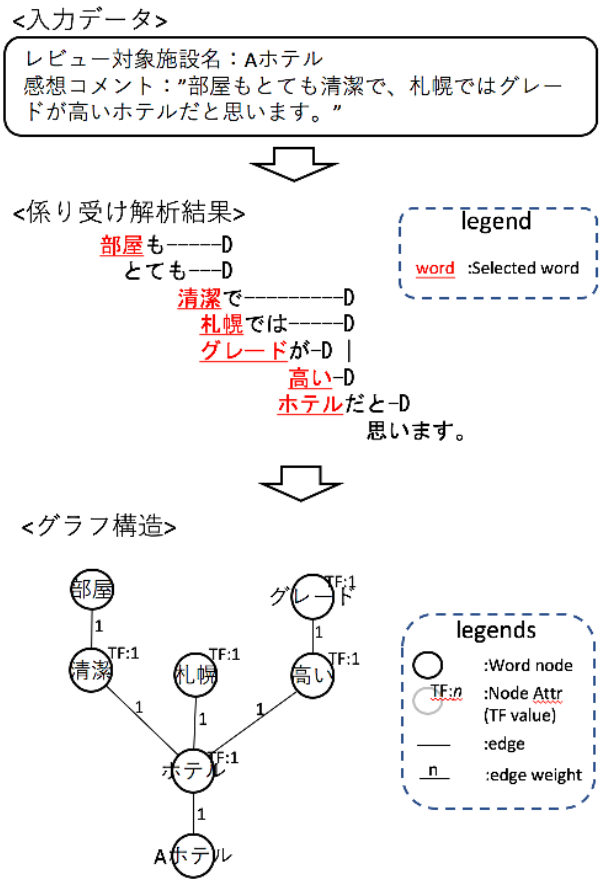

Fig.2 係り受け解析結果とグラフ作成

Fig.2 は感想コメントの係り受け解析結果と作成されるグラフ構造を示している。なお、係り受け解析結果は Cabocha のデフォルト出力形式である簡易 Tree 形式で表されている。

係り受け解析結果からグラフ構造を作成する際には、まず、特定の品詞の語を抽出する。Ⅱ.A で述べたように，評価属性辞書においては、〈属性〉には”大浴場”や“ベッド”など具体的な評価対象物が設定され、〈評価〉には”広い”などの評価物の状態が設定される。つまり、評価対象物は名詞であり、評価物の状態は文中では修飾語として現れる。このため、ここでは、名詞、形容詞、副詞を選択することとした。

次に選択した語と係り受け構造からグラフデータを作成する。選択した語をノード（以下、単語ノード）とし、係り受け関係をエッジとする。係り受け先のない単語(Fig.3 では”ホテル”)は、評価レビュー対象施設名を表すノード（Fig.3 では”A ホテル”）に連結される。

この時、語の出現回数（Term Frequency）（以下、TF 値）を単語ノードのアトリビュートとして設定する。また、ある 2 単語間の係り受け関係が出現した回数をその 2 単語ノード間のエッジの重み(weight)として設定する。

Fig.2 は一つの感想コメントからグラフを作る例であるが、分析では複数の感想コメントに対して処理を行う。この際、Fig.3 に示すようなグラフのマージを行う。

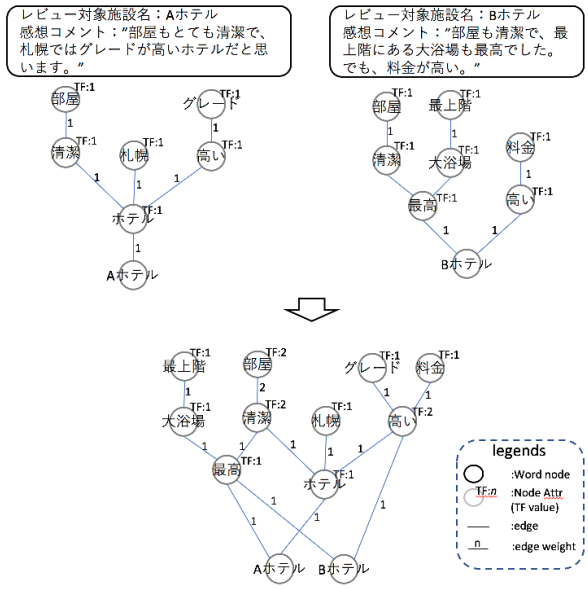

Fig.3 感想コメントグラフのマージ

Fig3.は Fig.2 で示した感想コメントのグラフに、別の感想コメントのグラフをマージする様子を示している。２つのグラフで同一のタンゴノードがあった場合

には、その単語ノードは共有される。この際、ノードアトリビュートの TF 値は合算される。Fig.3 では”部屋”,”清潔”,”高い”が両方の感想コメントで使われているため、マージ後のグラフではこれらの単語ノードの TF 値は２となっている。また、同一の２単語ノード間のエッジがあった場合にも、エッジの weight が合算される。Fig.3 では”部屋”-”清潔”の係り受け関係が両方の感想コメントで使われているためこれらの単語ノード間のエッジの weight は２となっている。

全ての感想コメントに対して上記の処理を行って作成したグラフに対して、Step.5 でグラフのクラスタリングを行う。

Step.5 のクラスタリングにおいては、R.Lambiotte らの手法[19]を用いた。この手法ではエッジの weight を考慮したクラスタリングが行われる。

Step.6 では、クラスタリングした結果から、クラスタ毎の特徴語を抽出する。具体的には、ノードのアトリビュートに設定した TF 値が高いノードを抽出する。

IV. 特徴表現抽出結果

Ⅲ章で述べた手法を用いて、宿泊レビューデータ中の感想コメントから、特徴表現を抽出した結果を以下に示す。

宿泊レビューデータは辻井らの研究[6]で利用されたデータと同じものを用いた。レビュー対象施設数はビジネスエリアの 50 施設、感想コメント数は 4986 件である。

まず、Step.4 で作成されたグラフデータの特徴を Table.1 に示す。

Table. 1 グラフ特徴量

ノード数	8302
エッジ数	56071
Average Degree（平均次数）	13.508
Diameter （直径）	9
Average Path length	3.217
Average Clustering Coefficient (平均クラスタ係数)	0.300

上記のグラフを Step.5 でクラスタリングした結果、17 のクラスタに分割された。

Fig.4 は各クラスタに含まれる単語ノード数と平均TF値に関係を示している。

Fig.4 に示すとおり、クラスターNo.0 から 12 は、単語ノード数と平均 TF 値共に比較的高く、クラスターNo.13 から 16 の単語ノード数と平均 TF 値は共に極端に低い。以下にクラスターNo.0 から 12 の内容について示す。なお、クラスターNo.13 から 16 の内容を確認したところ、特定の店舗やイベントに関する話題に関する単語が含まれていることがわかった。

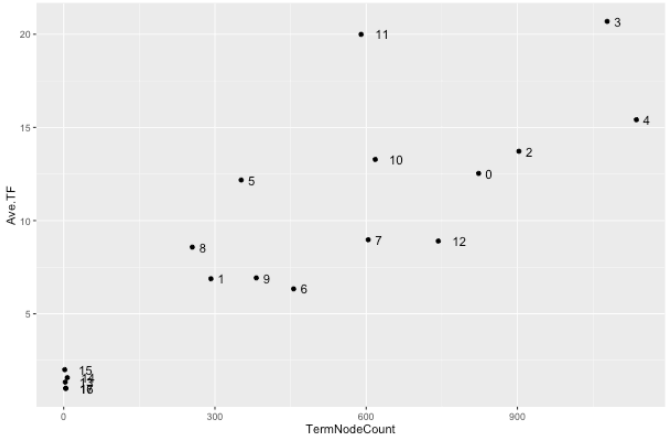

Fig.4 クラスタ毎の単語ノード数と平均 TF 値

Table.2 にクラスタ毎に TF 値の高い上位２０単語を示す。

Table.2 クラスタ毎の上位２０単語

#	Top 20 terms
0	便利,駅,近い,立地,場所,最高,コンビニ,食事,博多駅,繁華街,条件,地下鉄,車,大阪駅,移動,買い物,周辺,徒歩,直結,銀座
1	アメニティ,充実,タオル,空気清浄機,シャンプー,アメニティー,セット,香り,設置,強い,難点,備え付け,空気,AROMA,リセッシュ,交換,リンス,歯ブラシ,清浄,グッズ
2	朝食,満足,大変,美味しい,プラン,バイキング,多い,おいしい,特に,種類,無料,パン,レストラン,豊富,メニュー,少ない,ありがたい,和食,料理,店
3	部屋,広い,お部屋,風呂,ベッド,綺麗,狭い,快適,清潔感,清潔,ルーム,きれい,バス,夜景,トイレ,案内,十分,大きい,素敵,シャワー
4	利用,宿泊,出張,初めて,旅行,機会,名古屋,いつも,札幌,是非,大阪,子供,よく,観光,東京,ぜひ,一人,疲れ,福岡,2 人
5	値段,料金,安い,高い,駐車場,かなり,人,価格,得,リーズナブル,コストパフォーマンス,ポイント,割,金額,駐車,レベル,設定,正直,通常,追加
6	サービス,嬉しい,気分,感動,感激,結構,贅沢,帝国ホテル,ケーキ,ビックリ,心,サプライズ,本当にありがとうございました,新聞,欲,文句なし,浴衣,リッチ,堪能,個人的
7	チェックイン,予約,チェックアウト,荷物,お願い,遅い,早い,到着,スムーズ,確認,事前,12 時,台

	風,早め,急遽,おかげ,ちゃんと,満喫,夜中,じゃらん
8	エレベーター,安心,カード,フロア,必要,キー,混雑,セキュリティ,鍵,ラウンジ,配慮,クラブ,システム,扉,セキュリティー,エレベータ,エグゼクティブ,ロッカー,女性専用,防音
9	大浴場,温泉,うれしい,最上階,サウナ,露天風呂,地下,入浴,1 階,魅力,可能,窮屈,天然,脱衣,部分,24 時間,露天,表示,シティホテル,畳
10	ホテル,感じ,いい,他,新しい,雰囲気,ビジネスホテル,古い,普通,さすが,建物,口コミ,楽しい,客,自体,施設,評価,内装,おしゃれ,好き
11	良い,対応,フロント,よい,スタッフ,接客,丁寧,非常,親切,女性,素晴らしい,気持ちよい,すごい,本当に,アクセス,印象,気持ち,顔,方々,気持ち良い
12	ない,残念,無い,気,悪い,期待,臭い,問題,自分,不便,もう,不満,説明,水,まあ,一つ,申し分,仕方ない,足,タバコ

クラスタ番号 0(以下、#0 のように示す)では、”駅”,”近い”というように駅の立地に関する語が見られる他、”食事””繁華街””買い物”といった行動の目的となる単語も出ている。このクラスタは、宿泊施設のアクセスに関する特徴語が集まっている。#1 では、”アメニティ””充実”がトップであり、タオル,空気清浄機,シャンプーなど具体的な名詞がならんでおり、アメニティに関する特徴語が集まっている。なお、#1 の単語中に Amenities が２回出現しており、他のクラスタでの単語の重複がみられるが、これは Fig. 5 に示すように、実際の日本語単語では表記が異なるためである。長音記号の有無や、漢字表記/ひらがな表記などの表記揺れに加え、日本語/英語のカタカナ表記といった同じ意味の語が同じクラスタに現れていることが確認できた。

Amenities	By all means	Room	surprise
アメニティ	是非	部屋	びっくり
アメニティー	ぜひ	お部屋	サプライズ

Fig.5 日本語の表記揺れ

#2 は”朝食””満足””美味しい””バイキング””メニュー”など食事に関する語、#3 は”部屋””風呂””ベッド””広い””綺麗””清潔感”など、部屋に関する語が集まっている。また、#5 では”値段””安い””高い””コストパフォーマンス”といった価格に関する語が集まっており、#7 では”チェックイン””予約””チェックアウト””遅い””早い”といったチェックイン・チェックアウトに関する語があつまっている。さらに、#9 では”大浴場”,”温泉”,”サウナ”,”露天風呂”など浴場・温泉に関する語が、#11 では”良い”,”対応”,”フロント”,”スタッフ”,”接客”などスタッフの対応に関する語が抽出されている。

#4 では、”利用””宿泊”という語に続き、”出張”旅行””機会” “観光”という語や地名、”ひとり””子供”などが出現している。これは、誰とどのような機会において利用したのかという、オケージョンに関する語が抽出されたものと思われる。

#6 では”サービス””嬉しい””気分””感動””感激”などといった語が見られる。これは、宿泊者が体験した嬉しい出来事について語ったものがであり、カスタマーエクスペリエンスに関わる語が抽出されたものと思われる。

#8 では”カード”キー”がないと利用できない”エレベータ”や”ロッカー”など”セキュリティー”に関するものや”女性専用”や”防音”といった安全・安心に関する語が抽出されているものと思われる。

#10 では,”ホテル””ビジネスホテル”と共に”いい””雰囲気””古い”など、宿泊施設全体評価に関する語が抽出されたものと思われる。

#12 では、”ない”,”残念””悪い””問題”“不便”と共に”臭い””水””タバコ”などの語が見られ、不満点に関する語が抽出されたものと思われる。

上記で述べた各クラスタに含まれる語から推測される各クラスタが表す話題を Table.3 に示す。

Table.3 各クラスタの推測される話題

#	話題
0	アクセス
1	アメニティ
2	食事
3	部屋
4	オケージョン
5	価格
6	カスタマーエクスペリエンス
7	チェックイン チェックアウト
8	安心・安全
9	浴場・温泉
10	全体評価
11	スタッフ対応
12	不満点

辻井らの研究[6]では、評判分析の<属性>として、部屋・風呂・料理・接客・立地を挙げている。これらは Table3 では、それぞれ、#3,#9,#2,#11,#0 に対応する。また、辻井らは<属性>として”清潔感”も挙げているが、Table.2 に示したとおり、#3 の部屋に関する語とともに清潔感に関する語も抽出されている。また、田邉らの研究[4]では、宿泊レビューを分類・整理する際、辻井らと共通する、部屋，食事，風呂，接客・サ

ービスとい観点に加え、オケージョンという観点も用いている。これは Table3 では#4 に対応する。
つまり、Ⅱ.A で述べたように，宿泊レビュー分析では、評価対象となる<対象>とそれを細分化した〈属性〉〈評価〉が必要であったが、Table3 は<対象>に相当し、Table2 が〈属性〉〈評価〉に相当する。なお、前述の通り、〈属性〉は名詞であり、〈評価〉は修飾語である。このため、Table2 では、品詞を分けずに記載したが、品詞によって〈属性〉と〈評価〉に分けることが可能である。また、グラフ構造のデータを用いているため、どの〈属性〉がどの〈評価〉と関連しているかも容易に抽出できる。

また、Table.3 に示した通り、#2 アメニティ、#7 チェックイン・チェックアウト、#8 安心・安全など、辻井ら、田邉らの研究では分析観点として設定されていなかった観点も抽出された。

V. CONCLUSION

宿泊レビューに含まれる宿泊コメントの分析においては、これまで類義語辞書の作成が必要であった。類義語辞書の作成には手間がかかる他、随時見直しが必要になるといった問題や、想定していない評価対象に関しては類義語辞書が作成されず、想定外の評価対象に関しての評価が困難になると言った問題があった。

そこで、本研究では、語の共起制限と係り受け構造を利用して、感想コメント群をグラフ化してクラスタリングすることで、特徴表現を抽出した。これにより、予め評価対象を設定することや類義語辞書を作成することなく、宿泊レビュー中の感想コメントから評価に関する特徴表現を抽出することができた。

本研究では、ビジネスエリアの宿泊施設のレビューデータを用いたが、今後はレジャーエリアの宿泊施設のレビューデータなど記載内容の特徴が異なると思われるレビューデータで、本稿で述べた手法を適用し評価していく予定である。また、各クラスタが表す話題を含まれる語から推定したが、今後は推定の自動化を検討するつもりである。例えば、ノードの中心性などの指標を活用することが考えられる。

REFERENCES

[1] [JATA 2013]“数字が語る旅行業 2013”，日本旅行業協会 2013

[2] [Nielsen 2012] “Consumer Trust in Online, Social and Mobile Advertising Grows”, Nielsen,2012 http://www.nielsen.com/us/en/newswire/2012/consumer-trust-in-online-social-and-mobile-advertising-grows.html（2013/11/25 accessed）

[3] Chen, Yubo, and Jinhong Xie. "Online consumer review: Word-of-mouth as a new element of marketing communication mix." *Management science* 54.3 (2008): 477-491.

[4] 田湧亘. and 後藤正幸. "宿泊施設の戦略構築を支援するユーザレビュー分析に関する一考察 " *武蔵工業大学環境情報学部, 情報メディアセンタージャーナル* 9 (2008): 91-101..

[5] Tsujii, Koichi, Masakazu Takahashi, and Kazuhiko Tsuda. "Feature Extraction from Numerical Evaluation in Online Hotel Reviews." *Procedia Computer Science* 60 (2015): 1138-1145.

[6] 辻井康一, and 津田和彦. "テキストマイニングを用いた宿泊レビューからの注目情報抽出方法." *デジタルプラクティス* 3.4 (2012): 289-296.

[7] 小林 のぞみ，飯田 龍，乾 健太郎，松本 裕治,照応解析手法を利用した属性-評価対および意見性情報の抽出, 言語処理学会第 11 回年次大会, C2-6 (2005).

[8] Schutze, H. and Pedersen, J.O.: A Cooccurrence-based Thesaurus and Two Applications to Information Retrieval, International Journal of Information Processing and Management, Vol.33, No.3, pp.307–318 (1997).

[9] Miller, G.A.: WordNet: A Lexical Database for English, Comm. ACM, Vol.38, No.11, pp.39–41 (1995).

[10] Crouch, C.J.: A Cluster Based Approach to Thesaurus Construction, Proc. ACM SIGIR, pp.309–320 (1988).

[11] Chen, Z., Liu, S., Wenyin, L., Pu, G. and Ma, W.Y.: Building a Web Thesaurus from Web Link Structure, Proc. ACM SIGIR, pp.48–55 (2003).

[12] 中山浩太郎, 原隆浩, and 西尾章治郎. "Wikipedia マイニングによるシソーラス辞書の構築手法." *情報処理学会論文誌* 47.10 (2006): 2917-2928.

[13] 村上明子 那須川哲哉 “複数の筆者の表記の違いを利用した同義語抽出の精度向上” 言語処理学会第 8 回年次大会発表論文集 A1-5, 言語処理学会,3 月 2002

[14] 中渡瀬秀一 複合語からの類義語抽出法 情報 処理学会デジタル・ドキュメント研究会,Vol.32,No.6,pp.39046, 3 月 2002

[15] 卜野友司 et al. "係り受けの 2 部グラフと共起関係を利用した同義表現抽出." *情報処理学会研究報告自然言語処理 (NL)* 2004.1 (2003-NL-159) (2004): 169-176.

[16] Taku Kudo, Kaoru Yamamoto, Yuji Matsumoto: Applying Conditional Random Fields to Japanese Morphological Analysis, Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP-2004), pp.230-237 (2004.)

[17] mecab-ipadic-NEologd : Neologism dictionary for MeCab, https://github.com/neologd/mecab-ipadic-neologd

[18] 工藤拓, 松本裕治. チャンキングの段階適用による日本 語係り受け解析. 情報処理学会論文誌, Vol. 43, No. 6, pp. 1834–1842, 2002.

[19] Lambiotte, Renaud, J-C. Delvenne, and Mauricio Barahona. "Laplacian dynamics and multiscale modular structure in networks." *arXiv preprint arXiv:0812.1770* (2008).

[20] .