比較文集約に基づく主観的評価における補間エンティティの発見

旭　直人† 山本　岳洋†,†† 中村　聡史† 田中　克己†

† 京都大学大学院情報学研究科社会情報学専攻　〒 606–8501 京都府京都市左京区吉田本町

†† 日本学術振興会特別研究員（DC1）

E-mail: †{n.asahi,tyamamot,nakamura,tanaka}@dl.kuis.kyoto-u.ac.jp

あらまし　本稿では，比較文のマイニングによってエンティティ間の順序関係を明らかにし，ある観点で見た場合に２つのものの間に当てはまるエンティティ（補間エンティティ）及びその系列の発見を目的とする．間のものを発見するという意図の中でも特に，“この店よりはおいしくて，あの店よりはリーズナブルな店を見つけたい”，“2 つの本の間にあてはまるような難しさをもつ本を見つけたい”，といったような主観的評価における補間エンティティを発見する手法を提案する．提案手法では，検索エンジンによって比較文を発見，集約し，(評価対象，比較対象，評価，極性) の組を抽出し，それに基づきグラフを生成する．そしてグラフより補間エンティティ及びその系列を発見する．本稿では 3 つの発見手法を提案し，評価を行った．

キーワード　補間エンティティ，比較文集約

1. はじめに

近年，検索エンジンは必要とする情報を得るうえで欠かせないツールとなってきている．検索エンジンの進歩に伴い，ユーザが明確な検索意図を持っており，的確なクエリを作ることができれば，ユーザは自分の求める情報へ容易にたどり着くことができるようになった．しかし依然として，明確な検索意図を持っているにも関わらず，的確なクエリを思いつくことができなければ，求める情報へたどり着くことは困難である．このような状況はさまざま考えられるが，その中でもクエリをキーワードとして表現することが困難である状況が存在する．例えば以下のような場合がある．

- 戦国時代の出来事といえば，“桶狭間の戦い” と “本能寺の変” があった．しかし，その間に何があったかを思い出すことができない．この 2 つの出来事の間にあった出来事を知りたい．
- 世界最長の川といえばナイル川であることは覚えており，黄河も相当長いことを覚えている．そこで，長さの観点でナイル川と黄河の間にくるような川は何か調べたい．
- 両親を夕食に誘いたい．大学の近くにはたくさん飲食店がある．○○レストランでは安いがまずくて両親は満足しないだろう．しかし，××レストランだとおいしいが高すぎる．前者よりはいいが，後者よりは控えめといったようなレストランを見つけたい．
- 友達に宮部みゆきの本を 5 冊貸してもらい，最初と最後に読むべき本を勧めてもらった．その 2 冊の間だと，どういう順番で読むのがベストか知りたい．

これらの状況では，ユーザはすでに分かっている 2 つのエンティティの間にあたる何かを見つけたいと考えている．このような検索は補完検索の一種であると考えられる．しかし，従来の検索エンジンではこうした情報を直接的に得ることは困難である．現在の検索エンジンは，文書が指定されたキーワードを含むかどうかという基準で結果を返すが，この状況においてユーザは知りたいものの名前をそもそも知らないため，直接その名前を用いてクエリを生成できない．結果として，ユーザは欲しい情報に直接到達できない．また，適当なクエリが得られ，情報が載っているページへ辿りつけても，今度はページ内から必要な情報を探し出し，判断する必要がある．そのため，このような情報を得るためにはクエリを工夫したり，似たようなクエリを何度も作り，多くのページを閲覧する必要がある．

この問題を解決するため，我々はこれまでに既知の 2 つのエンティティをクエリとして受け取り，ある観点でエンティティを順序づけた場合に，その 2 つのエンティティの間に存在するようなエンティティ（本研究ではこれを補間エンティティと呼ぶ）をユーザに提示する検索手法を提案してきた．ある観点における補間エンティティを発見することは，エンティティ検索において特定の観点での順序を考慮した検索で有用であると考えられる．ここでいうある観点とは，先の例であれば，時間，長さ，品格，おもしろさといったものになる．これらの観点は客観的なものと主観的なものに分けることができる．例えば，先述の 4 つの例の内，最初の 2 つの例の場合，時間や長さという観点であり，こうしたものは客観的事実であるため，誰が並べても同じ順番が得られる．一方，後の 2 つの例の場合，品格やおもしろさといった観点であるため，人によって並べる順番が変わるような主観的なものであると言える．つまりこのように観点は客観的評価と主観的評価に基づくものに分類することが出来る．我々はこれまで客観的事実における補間エンティティの抽出に取り組んできた [1] [2]．しかし，その手法では発見したい順序と文書上での記述順序が一致するという仮定をおいていたため，求めたい順序と記述する順序がほとんど一致しない “おもしろさ” や “難しさ” といったような主観的な評価における補間エンティティを発見することができなかった．

そこで本稿では，主に主観的評価における補間エンティティの抽出に取り組む．提案手法では，Web から比較を表す文を取得し，比較文に含まれる比較関係からエンティティ間の順序を推測する．比較文では客観的事実に基づく比較に限らず，2 エンティティに対する主観的な評価が記述されている場合もあり，これらを集約することによって，主観的評価におけるエンティティ間の順序を発見することが可能であると考えられる．

実際には，まず与えられた 2 エンティティと比較されうる語集合である比較対象集合（与えられた 2 エンティティを含む）を与える．そして，得られた比較対象集合に含まれる語と比較文に特有な構文パターン（「～より」「～のほうが」など）を組み合わせてクエリを生成し，Web からの比較文の抽出を行う．最後に，得られた比較文から (評価対象，比較対象，評価，極性) の組を比較文から抽出し，それらを集約することによって，ある観点におけるエンティティ間の評価を表すグラフを生成する．そして，生成したグラフから与えた 2 点間で最適なパスを発見することで，補間エンティティ及びその系列を発見する．本稿では推移律を考慮した手法，枝の重みを重視した手法，枝の類似度を重視した手法の 3 つを提案し，それぞれの手法に関して実験，評価を行った．

2. 関 連 研 究

比較は一般的に同位語関係にある語同士で行われる．同位語とは，“トヨタ” や “日産” に対する “ホンダ”，“ダイハツ” といった共通の上位語をもつ語のことである．比較対象集合に含まれる語は同位語関係になる可能性が高い．この同位語を発見する研究について多数行われている．Ghahramani ら [3] の Bayesian Sets では，語の共起テーブルのような大規模なデータに対し，ベイズ推定を用いることで同位語を取得する．我々のグループ [4] は，検索時において同位語同士では似たようなクエリが用いられることに注目し，クエリログから同位語を発見する手法を提案している．Lin ら [5] の提案では，係り受け解析が行われている大規模コーパスを用いて，類似する語のクラスタを発見する．Shinzato ら [6] の提案では，HTML 構造を用いて同レベルに記述されている語を同位語の候補として取得し，相互情報量，共起度が高いものを同位語として評価する．我々のグループ [7] は同位語が記述される際の助詞に着目し，検索エンジンの返すタイトルやスニペットを利用して巨大なコーパスを持たずとも同位語のコンテキストを考慮しながら同位語を発見する手法を提案している．Talukdar ら [12] は自動的にトリガーとなる言葉を発見し，ブートストラップ的手法で抽出パターンを発見する手法を提案している．Wang ら [9] は，同位語は例えば “< li >～” といったような同じような記述に囲まれている事が多いという性質を利用し，複数の Web ページから同位語を発見する手法を提案している．Web から同位語を発見する研究には，KnowItNow [10] [11] といったような研究もある．Web サービスとしては，複数の語を与えることで同位語の集合を返す Google Sets [8] がある．本稿では，比較文に現れる構文パターンを利用して比較対象集合の拡張を行っている．

また，商品のレビュー記事などから主観的な評価をマイニングする研究とも本研究は関連が深い．Liu ら [13] はポジティブ，ネガティブが分離されて記述されているようなレビュー記事から商品の属性とそれに対する評価を抽出し，商品に対する評価を可視化する手法を提案している．Liu らの手法は個々の商品に対しての評価を求めており，他の商品との比較には基づいていない．

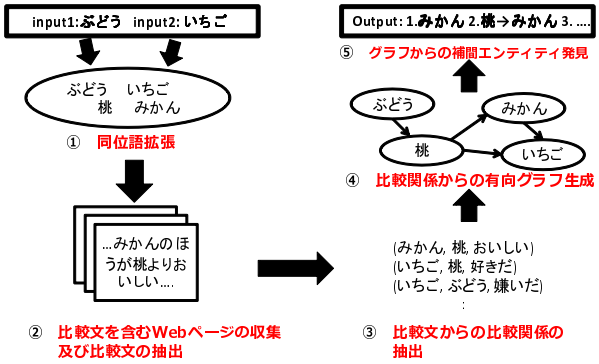

図 1 手法の流れ

比較文取得に関する研究も盛んに行われている．Jindal ら [14] [15] は，言語パターンに注目し，Class Sequential Rules とナイーブベイズ分類器を用いて文書中から比較文を発見し，比較関係の抽出も行っている．Kurashima らは [16] ブログから言語パターンを用いて比較関係を抽出し，リンク解析をすることでエンティティ間のランキングを行っている．比較文抽出手法に関して，本稿での提案手法で参考にしている．佐藤ら [17] は，言語パターン，センタリング理論を用いることでブログから比較表現，比較関係を抽出している．

3. Web からの比較文マイニングとエンティティ間のランキング

3.1 概 要

本章では，Web から比較文をマイニングし，エンティティ間の順序を明らかにする手法について述べる．

比較文とは，「プラズマテレビの輝度は，液晶テレビと比べて，おおよそ半分程度だ」，「コーヒーより紅茶のほうが好きだ」といったような 2 つのものをある観点での比較を述べている文である．本稿では，比較文を構成する要素として，比較関係（評価対象，比較対象，評価） の組を考える．“比較対象” を基準として，“評価対象” がある観点において評価される．例えば，先ほどの例だと評価対象は “プラズマテレビ”，“紅茶”，比較対象が “液晶テレビ”，“コーヒー” になり，評価はそれぞれ “半分程度である”，“好きだ” になる．また，各評価はポジティブとネガティブの極性をもつ．“半分程度である” だと，ネガティブな意味になり，“好きだ” はポジティブな意味になる．評価対象及び比較対象は必ずしも比較文に明示的に記されるわけではなく，省略されている場合がある．例えば，「· · · Android 携帯に興味がある．しかし，iPhone と比べると UI は洗練されていない．· · ·」といった文章の場合，評価対象は “Android 携帯” であるが比較文では省略されている．また，厳密には評価は評価対象，比較対象のある属性に対して行われる．プラズマテレビと液晶

表 1　作成するクエリ（$t \in C$）

t は比較対象	“t と比べ”
	“t に比べ”
	“t と比較”
	“t より”
t は評価対象	“t のほうが”
	“t の方が”

テレビの例の場合だと，それぞれの輝度について言及しており，コーヒーと紅茶の場合，それぞれの味に関して評価が行われている．本稿では簡単のため，エンティティの属性までは考慮に入れない．

提案手法では，以下のようにして比較文のマイニングとエンティティ間のランキングを行う（図 1）．

（1） 補間エンティティを求めたい 2 エンティティを入力として与える
（2） 得られた入力と比較の対象になりうる語の集合（比較対象集合）を求める
（3） 得られた語集合と比較文特有の表現を用いて，Web から比較文のマイニングを行う
（4） 得られた比較文に含まれる評価対象，比較対象，評価の極性を求め，極性に基づいたグラフを生成する
（5） グラフより 2 点間のパスを求め，得られたパスのランキングを行い，補間エンティティ及び系列を提示する

本稿では，(2) の入力からの比較の対象となる集合は手動で与える．なお，比較対象集合獲得の自動化については 5. 章で述べる．以下の節で，(3)，(4) のそれぞれについて詳細に述べる．

3.2　Web からの比較文マイニング

得られた比較対象集合 C を用いて，Web からの比較文の抽出を行う．比較文では，「〜より」「〜と比べ」「〜のほうが」といったような特有の構文が用いられる．これらの比較文に典型的な構文と比較対象集合に含まれるエンティティを組み合わせてクエリを作成する．本稿では表 1 のように 1 つのエンティティに対し，6 種類のクエリを作成した．「t より」「t と比べ」「t に比べ」「t と比較」のクエリの場合，t は比較対象となり，「t の (ほう | 方) が」のクエリの場合，t は評価対象となる．Yahoo Web Search API (注1) を用い，それぞれのクエリに対して，上位 100 件を取得し，それぞれのウェブページをダウンロードする．そしてそれぞれのページにおいて，そのページを取得したクエリの正規表現を用いて，比較を表している文を抽出する．この時，クエリに含まれている t は確実にその比較文に含まれているが，t に対する評価対象及び比較対象が抽出した文に含まれているとは限らない．そこで，抽出した比較文に評価対象及び比較対象のいずれかが発見できなかった場合は，t に対する評価対象及び比較対象は比較文の直前に現れる可能性が高いと仮定し，比較文より前の文に現れている要素 $t' \in C$（ただし，$t' \neq t$）を評価対象及び比較対象として扱う．例えば，図 2 の場合では，比較文では比較対象が省略されてしまっている．そこで，比較文より前を探索し，そこに比較対象となる “Perl” を発見したので，“Perl” を比較対象として扱う．

(注1)：http://developer.yahoo.co.jp/

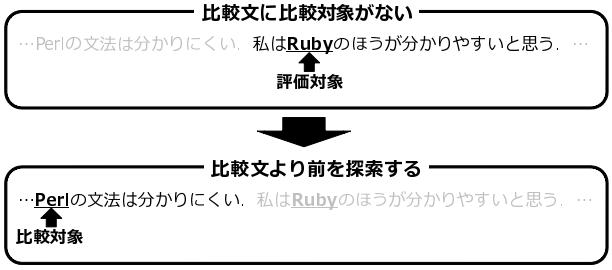

図 2　評価対象及び比較対象が比較文で省略されていた場合

表 2　観点 “おもしろい” に対応する辞書

評価語	極性
美味しい	ポジティブ
おいしい	ポジティブ
良い	ポジティブ
うまい	ポジティブ
美味い	ポジティブ
旨い	ポジティブ
まずい	ネガティブ
不味い	ネガティブ

次に，比較文に含まれる評価について説明する．評価は「難しい」「おいしい」「きれいだ」といったような形容詞や形容動詞で表されている場合もあれば，「興味ある」「目立つ」といったように動詞で表されている場合，「逸材」「難攻不落」のように名詞で表される場合もある．また，これらはポジティブ，ネガティブの 2 つの極性に分類することが出来る．本稿では，特定の観点における評価のみを抽出するために，それぞれの観点ごとに手動で辞書を用意した．辞書には，それぞれの観点に対応する評価を表す語とそれに対応する極性（ポジティブ，ネガティブのいずれか）を記述している．例えば，「おもしろい」という観点に対応する辞書は表 2 のようになる．この辞書とのマッチングによって，比較文に含まれている求めたい観点における語句の極性の判定を行う．これにより，比較関係（評価対象，比較対象，極性）の組を得ることが出来る．比較関係（評価対象，比較対象，極性）の組の集合を，$R = \{(v, w, p)\}$ とし，v を評価対象，w を比較対象，p は *positive* または *negative* のいずれかをとるとする．

3.3　比較関係グラフの構築

得られた比較関係集合 R を用いて，C に含まれるエンティティをノードとした有向グラフ $G = (V, E)$ を作成する．ノード集合 V は $V = C$，枝集合 E は

$$E = \{(v, w) | ((w, v, positive) \vee (v, w, negative) \in R\} \quad (1)$$

となる．つまり，比較文に含まれている評価がポジティブの場合，比較対象を表すノードから評価対象を表すノードへの有向枝を張る．逆に，比較文に含まれている評価がネガティブの場合，評価対象を表すノードから比較対象を表すノードへの有向枝を張る．ノード v から w に向けての枝が張られている場合，何らかの観点で v より w のほうが優っていることを表している．$sf(v, w, p)$ を (v, w, p) を満たす比較文の数であるとすると，

ノード v から w に向けての枝の重み $weight(v,w)$ は，

$$weight(v,w) = sf(w,v,positive) + sf(v,w,negative) \qquad (2)$$

と表される．

次に，枝の特徴ベクトルを定義する．これは比較文は様々なコンテキストで述べられているため，複数の比較文を集約する際にはなるべく同じようなコンテキストであるようにすることが必要だと考えられるからである．まず，各比較文において，比較文とその前後の 25 文字をあわせた文字列を *MeCab* (注2) によって形態素解析し，名詞と名詞句，動詞を抽出し，各単語の出現頻度で特徴ベクトルを作成する．枝 (v,w) を構成する k 番目の比較文の特徴ベクトルを $tf_k(v,w)$ とした時，枝 (v,w) の特徴ベクトル $vector(v,w)$ を以下のように定義する．

$$vector(v,w) = \sum_{k=1}^{weight(v,w)} tf_k(v,w) \qquad (3)$$

このように，各枝に重み及び特徴ベクトルを付与したグラフ $G=(V,E)$ を生成する．

3.4 補間エンティティ及びその系列の発見

次に生成したグラフから補間エンティティを発見する手法について説明する．ユーザは挟みこみたい 2 つのエンティティを入力する際に，どちらがある観点で優れているかということを含めて入力する．例えば，長さという観点でナイル川と黄河を与える際に，ユーザはナイル川のほうが黄河より長いということを指定して入力する．つまり，グラフの中での始点と終点は分かっているという状態となる．よって，グラフより補間エンティティを見つけるということは 2 点間の最適な経路を求めることに他ならない．経路の長さが 3（経路が始点 $\rightarrow s_k \rightarrow$ 終点となる場合）の時，出力されるものはノード s_k となり，これは補間エンティティにあてはまる．一方で経路の長さが 4 以上（経路が始点 $\rightarrow s_k \rightarrow \cdots \rightarrow s_{k+l} \rightarrow$ 終点となる場合）の場合，出力されるものは始点と終点を除いた 2 点間の経路 $s_k \rightarrow \cdots \rightarrow s_{k+l}$ となる．この場合，補間エンティティの系列となる．本稿では，枝の重みに着目した手法，枝の特徴ベクトルに着目した手法，推移律に着目した手法の 3 つの手法を提案する．

3.4.1 重み重視手法

多くの比較文で言及されている関係，つまり重みの大きい枝は信頼できるのではないか，と仮定する．これは多くの人に支持されている関係のほうが信頼できるという考えから来ている．加えて，経路は短いほうが信頼できると仮定する．これは，枝の基となる R は様々な人の意見から構成されているため，経路が長くなればなるほど様々なコンテキストを含みやすくなり，経路の信頼性が下がると考えられるからである．これらの仮定に則り，始点から終点までに通るノードの数を可能な限り少なくし，かつ，重みの大きい枝を通るような経路を探索する．これは，経路に含まれる枝の重みの逆数の和を最小にする問題と言い換えることが出来る．ただし，生成されたグラフ $G=(V,E)$ は閉路を含むため，経路を探索する際には，既に通ったノードは二度通らないものとする．始点 s_1 から終点 s_n へのある経路 $path_k$ を $path_k = \langle (s_1,s_2),(s_2,s_3),\cdots,(s_{n-1},s_n)\rangle$ とした時，$path_k$ のスコア $Rank_{weight}(path_k)$ を次のように定める．

$$Rank_{weight}(path_k) = \sum_{i=1}^{n-1} \frac{1}{weight(s_i,s_{i+1})} \qquad (4)$$

この $Rank_{weight}$ の値が小さい経路から上位にランキングしていく．

3.4.2 類似度重視手法

重み重視手法では枝のコンテキストについてはほとんど考慮出来ていない．そこで，経路に含まれる枝のコンテキストが似ていれば似ているほど，その経路の一貫性は保たれると仮定し，枝の特徴ベクトル $vector(v,w)$ を利用した手法を考える．枝の特徴ベクトル $vector(v,w)$ はその枝のコンテキストと考えることができる．よって，経路に含まれる枝のコンテキストが一致しているということは，枝同士の特徴ベクトルの類似度が高いのではないか，と考えられる．そこで，特徴ベクトル v_1，v_2 のコサイン類似度を $Sim(v_1,v_2)$ とした時，経路 $path_k$ に対する類似度を用いたスコア $Rank_{sim}(path_k)$ を次のように定める．

$$Rank_{sim}(path_k) = -\sum_{i=2}^{n-1} log\left\{Sim\left(\sum_{j=1}^{i-1} vector(s_j,s_{j+1}), vector(s_i,s_{i+1})\right)\right\} \qquad (5)$$

ただし，$Sim\left(\sum_{j=1}^{i-1} vector(s_j,s_{j+1}), vector(s_i,s_{i+1})\right) = 0$ となる時は，$log\left\{Sim\left(\sum_{j=1}^{i-1} vector(s_j,s_{j+1}), vector(s_i,s_{i+1})\right)\right\} = -100$ とする．今まで通ってきた経路の特徴ベクトルの和と次に進む枝の特徴ベクトルとの類似度を足していったものをこの式は表している．この $Rank_{sim}$ の値が小さい経路から上位にランキングしていく．

3.4.3 推移律を考慮した手法

グラフ $G=(V,E)$ から得られる順序は，半順序である．半順序とは，反射律，反対称律，推移律を満たすものである．ここで，推移律に注目する．推移律とは，$a \leqq b$，$b \leqq c$ ならば $a \leqq c$ が成り立つことをいう．重み重視手法と類似度重視手法では，推移律のことは考慮出来ていなかった．本節では，推移律を考慮した手法について説明する．

図 3 のように経路 $\langle (A,B),(B,C)\rangle$ が存在するとき，(A,C) はこの経路の推移律を満たす．よって，(A,C) は経路 $\langle (A,B),(B,C)\rangle$ を支持する要素となり得る．そこで，(A,C) の重み $weight(A,C)$ を $weight(A,B)$ と $weight(B,C)$ に配分することを考える．元の $weight(A,B)$ と $weight(B,C)$ の関係をある程度維持するため，$weight(A,B)$ と $weight(B,C)$ の比に応じて $weight(A,C)$ を分配する．経路の長さが長くなった場合も同様にして，経路に対する全ての推移律を満たす枝の重みを経路に含まれる枝に分配していく．具体的にはある経路 $path_t$ の長さを n としたときに，図 4 のようなアルゴリズムを用いて，経路上の各枝の新しい重み $weight'(v,w)$ を計算する．そして，新たに得られた枝の重み $weight'(v,w)$ を用いて，

（注2）：http://mecab.sourceforge.net/

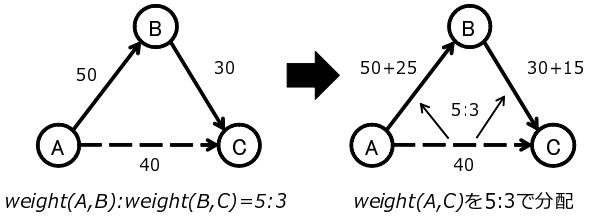

図 3　3 点の場合の推移律を考慮した枝の重み配分

```
for(i = 1; i < n - 1; i++){
  for(j = 2; i + j < n + 1; j++){
    base = weight(i, i + j)
    sum = 0
    for(k = i; k < i + j; k++){
      sum += weight(k, k + 1)
    }
    for(k = i; k < i + j; k++){
      if(weight'(k, k + 1) == 0){
        weight'(k, k + 1) = weight(k, k + 1)
      }
      weight'(k, k + 1) += base * weight(k, k + 1) / sum
    }
  }
}
```

図 4　推移律を考慮した枝の重みの再計算

$$Rank_{sliding}(path_t) = \sum_{i=1}^{n-1} \frac{1}{weight'(s_i, s_{i+1})} \tag{6}$$

を求める．このようにグラフから得られる全ての経路に対して再計算を行い，$Rank_{sliding}(path_t)$ の小さいものから順に提示する．

4. 実　　験

提案手法の精度を評価するために実験を行った．実験はユーザアンケートに基づくものと一般的な回答に基づくものの 2 つ行った．ユーザアンケートに基づく実験では，6 人の被験者にアンケートをとって，各テストセットで順位をつけてもらい，そのばらつきの違いによって結果が各手法でどのように変化するかを観察した．一般的な回答に基づく実験では，Web 上で行われている投票や，ある機関が行っている調査による結果に基づき，正解セットを作成し，どの程度の精度で抽出できるかを確認した．用意した辞書はそれぞれ好きな順に関する辞書には 8 語，値段の安い順に関する辞書には 3 語，難しさ順に関する辞書には 10 語登録されている．

4.1　ユーザアンケートに基づく精度評価

提案手法がどのくらい有効かを調べるために，まずユーザアンケートを行い，正解セットを作成した．6 人の被験者に対し，表 4 に示す 13 個のテストセットを用意し，それぞれに対して指定した観点で用意した要素を 1 位から順に順位付けを行ってもらった．その後，各テストセットに対してそれぞれのユーザが決めた順位から平均の順位を求め，それを正解の順序とした．次に，自動的に正解，不正解を判定するために，少なくとも 1 つ以上の正解となる補間エンティティ及び系列を含むようにして正解順序から 2 つの要素を全て組み合わせてクエリを生成し，それぞれのクエリで補間エンティティの取得をおこなった．例えば，正解順序が (A, B, C, D, E) であるとき，(A, C)，(A, D)，(A, E)，(B, D)，(B, E)，(C, E) というクエリを作成する．そして，抽出された経路が正解系列のクエリ間での順序関係を保っていれば正解，そうでなければ不正解と判定する．例えば，先の例でクエリが (A, E) のとき，B，B → C，B → D は正解と判定され，C → B，　B → D → C は不正解と判定される．評価尺度は平均逆数順位 (MRR) を用いる．MRR は以下の式（7）で表される．K はクエリ数，r_i は i 番目のクエリにおいて，適合する系列が出現した最上位の順位である．ただし，i 番目のクエリにおいて，適合する系列が存在しない場合は $\frac{1}{r_i} = 0$ とする．

表 4　アンケートで用いたテストセット

カテゴリ	並べる観点	要素数
コンビニ	好きな順	6
音楽家	好きな順	7
お菓子	好きな順	7
季節	好きな順	4
天気	好きな順	4
科目	好きな順	5
鉄道	好きな順	6
大学	難しい順	7
牛丼屋	美味しい順	4
牛丼屋	安い順	4
ハンバーガー屋	美味しい順	4
ハンバーガー屋	安い順	4
SMAP	好きな順	5

$$MRR = \frac{1}{K}\sum_{i=1}^{K}\frac{1}{r_i} \tag{7}$$

推移律を考慮した手法，重みを重視した手法，類似度を重視した手法のそれぞれ 3 つの場合で MRR を求めた．ベースライン手法として，それぞれのグラフ G から得られる全ての経路の中からランダムで経路を選択して並べるという手法を用いる．全経路の中での正解数を n_{ans}，全経路の数を N とすると，ランダム選択した場合の逆数順位の期待値 RR_{random} は

$$RR_{random} = \frac{n_{ans}}{N} + \sum_{k=2}^{N-n_{ans}+1}\left(\frac{1}{k}\cdot\frac{n_{ans}}{N-(k-1)}\prod_{l=0}^{k-2}\frac{N-n_{ans}-l}{n-l}\right) \tag{8}$$

となる．よって，K をクエリ数とすると，ランダム選択手法の MRR は

$$MRR_{random} = \frac{1}{K}RR_{random} \tag{9}$$

と表される．

結果は図 5 のようになった．比較文の抽出例は表 3 の通りである．平均で，推移律考慮手法が 0.49，重み重視手法が 0.47，類似度考慮手法が 0.47，ベースライン手法が 0.28 となった．全体的に推移律考慮手法が優れているが，類似度重視手法が優っている場合もある．推移律考慮手法が他の手法より優れていたものに，“SMAP” があるが，このグラフは図 6 のようになっている．これをみると，重みが少なくなっている経路があるが，推移律を考慮することによって，こういった重みが少なくなっているが正しい経路を上位にランキングすることができ，精度が良かったのではないかと考えられる．ここで，テストセット

表 3　得られた比較文の数及び例

	得られた比較文	辞書にマッチした比較文	実際に取得できた比較文の例
牛丼屋	1158	187	すき家よりやわらかくておいしかったです！ すき家よりも値段が多少安いので、気軽に鰻の味を楽しめる.
SMAP	1210	254	やっぱり中居よりつよぽんのが好き. 初めは慎吾のファンだったけど、最近、吾郎のほうが大好きになりました.
鉄道	1524	178	わたしは阪急のほうがすきだ. やはり電車的に京阪のほうがいい.

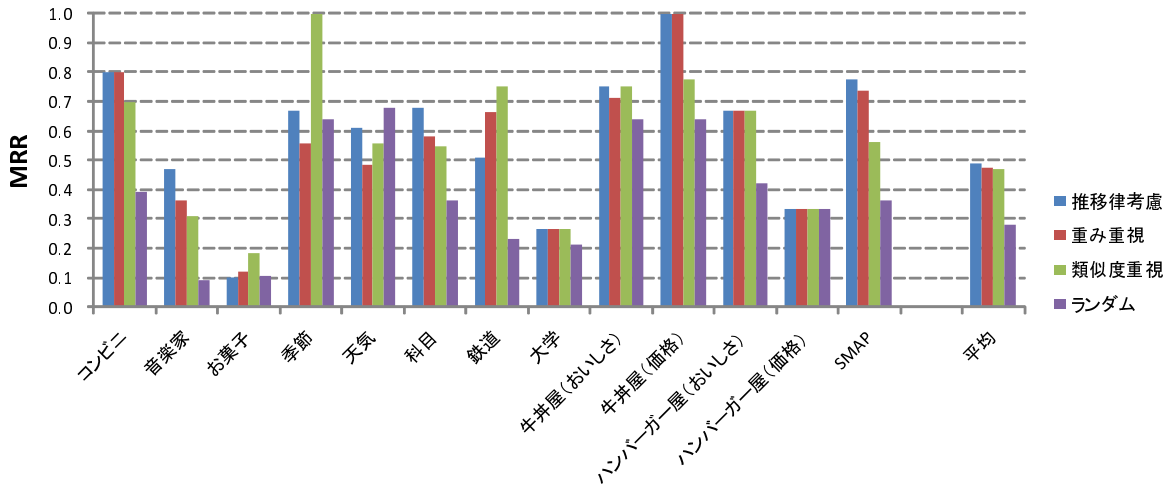

図 5　ユーザアンケートによる実験結果

をユーザ間で順序のつけ方にばらつきが出たものとそうでないものの 2 つのグループにわける．ばらつきが出たテストセットは “音楽家”，“お菓子”，“季節”，“科目”，“鉄道”，“SMAP” であり，ばらつきがあまり出なかったテストセットは “コンビニ”，“天気”，“大学”，“牛丼屋（おいしさ・価格）”，“ハンバーガー屋（おいしさ・価格）” であった．それぞれのセットにおける MRR は図 7 のようになった．ばらつきが出たグループでは，類似度重視手法の精度が高くなっている．これは，ばらつきが出るような場合には枝のコンテキストを重視するほうがよりよい結果を得られる可能性があると考えられる．ばらつきが出ないような場合には，推移律考慮手法が優れていた．

4.2　大規模アンケートに基づく精度評価

次に，ある機関が行っている大規模なアンケートによる調査や Web 上で行われているある事柄に関する投票は，一般の人の意見を集約したものであると考えられる．そこで本節では，それらの結果を一般的に認識されている順序と考え，正解順序として用いる．用いたテストセットのカテゴリと正解順序を定めるために利用した大規模アンケートは表 5 の通りである．前節と同様にそれぞれのテストセットで各手法の MRR を求めた．

その結果，図 8 のようになった．平均の MRR で見ると，推移律考慮手法が 0.65，重み重視手法が 0.63，類似度重視手法が 0.63，ベースライン手法が 0.27 となっており，提案手法はベースラインと比べ，精度が出ていると言える．推移律考慮手法は，全体的に精度が出ている．ユーザ毎に好みがばらつく傾向にあるテストセットには類似度重視手法が有効であることが前節の

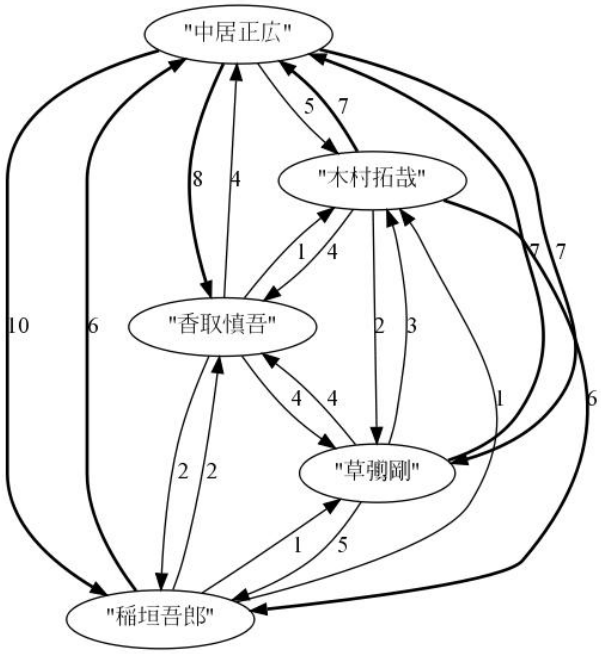

図 6　SMAP のメンバーで好きな順で生成されたグラフ

結果より，得られていたが，大規模アンケートに基づくテストセットに対しても，類似度重視手法が他の手法と比べ，良い精度を出しているテストセットが多いことがわかった．しかし，フルーツのテストセットで精度が大きく下がってしまっている．これは，ある番組で特定の組み合わせのフルーツが使われてい

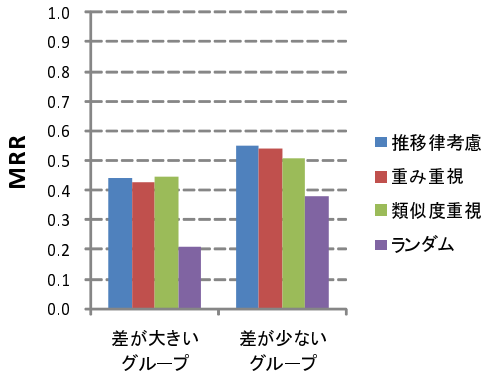

図 7　順位の付け方の差の大きさによる MRR の違い

表 5　用いたテストセットと利用した大規模アンケート

カテゴリ	要素数	利用したデータ
コンビニ	6	http://getnews.jp/archives/15676
音楽家	7	NHK 放送文化研究所世論調査部「日本人の好きなもの」2008 年
季節	4	NHK 放送文化研究所世論調査部「日本人の好きなもの」2008 年
科目	5	ベネッセ 第 4 回学習基本調査報告書・国内調査 中学生版
牛丼屋	4	http://getnews.jp/archives/81459
フルーツ	7	NHK 放送文化研究所世論調査部「日本人の好きなもの」2008 年
AKB48	5	AKB48 選抜総選挙結果

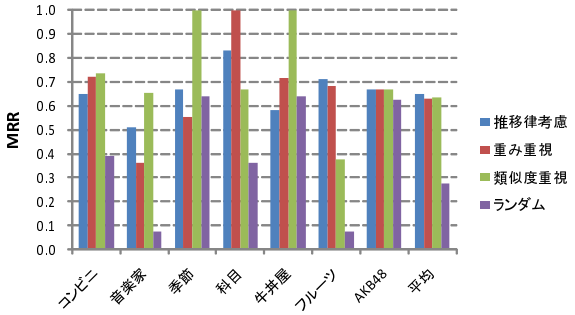

図 8　大規模アンケートによる実験結果

たため，そのフルーツを組み合わせた経路の類似度が高くなる傾向にあったが，それは一般的な好みと異なっていた．そのため，精度が下がったものと考えられる．このように，類似度手法は誤ったコンテキストが含まれていた場合，精度を大きく下げてしまうおそれがあることが分かった．

5. 考　　察

4.2 節の結果より，提案手法は大規模な調査・アンケート結果とある程度の精度で一致が見られた．これは，Web から比較文をマイニングすることにより，大規模な調査をせずとも，人々の主観的評価がある程度推測できる，ということであると考えられる．これは，なんらかのアンケート結果や表，ランキングがないような場合でも，ある観点における補間エンティティを抽出できることを示したと考えている．

次に各手法が有効に働く場合についてみてみると，4.1，4.2 節の結果より，人によって好みが分散するような場合には類似度重視手法が有効に働くと考えられる．そして，そうでないような場合には推移律考慮手法が有効である．このことより，好みの分散を考慮しながら推移律考慮手法と類似度重視手法を組み合わせることによって，全体的な精度の向上が見込めるのではないかと考えられる．また，今回は比較文から比較関係 (v, w, p) を抽出する際に，属性を考慮せず，評価対象または比較対象が比較文内に発見できなかった場合は，比較文より前の文を探索しに行き，最初に発見した $t \in C$ を評価対象または比較対象として扱うということを行った．このことは，比較関係 (v, w, p) の再現率を高める上では重要だが，適合率は下がってしまう．そのため，抽出ミスが含まれた状態でグラフの生成が行われてしまっている．比較関係の抽出精度を高めれば，グラフの正確さが増し，補間エンティティ発見手法の精度も上昇すると考えられる．現状では，特定の観点に限定した上で厳しく比較関係の抽出を行うと，数が集まらずグラフがうまく生成できなくなってしまう．今後，マイクロブログなどの Web サービスの充実により，Web 上により多くの比較文が記述され，検索技術の発達とともに，この問題は解決していくのではないかと考えている．

最後に，比較文の特性についての考察を行う．鈴木 [18] は「A は B より～」型の比較文について，「B は参照として機能していると考える．参照点とは，我々が，ある対象をとらえようとする際に手がかりとして用いる，よりとらえやすい，目標となる対象とは別の，認知的に際立った対象のことである．」と述べている．実際に，我々が音楽家に関して比較文を集約し，生成したグラフは図 9 のようになっており，“ベートーヴェン” や “モーツァルト” といった有名なノードに枝が集中しているのが分かる．そして，マイナーなノードにはあまり枝が集まらない傾向にある．推移律を考慮したとしても，そもそも枝が張られていなければ経路をたどることができないので，マイナーなノードを発見することは困難となる．このように，エンティティ間にメジャーなものとマイナーなものといった関係が存在すると，図 10 のようなメジャーなエンティティのノードに枝が集まり，マイナーなエンティティ間にはそもそも枝がほとんど張られないといった現象が起き，うまく補間エンティティを発見することが困難となる．よって，比較を行うエンティティ同士はなるべく対等な関係にあることが望ましい．

今後の課題として，2 入力からの比較対象集合拡大の自動取得が挙げられる．2 入力に対する比較対象集合に含まれうる語は，2 入力の同位語であると考えられる．そこで，Ohshima らの手法 [7] を用いて，自動拡張することが考えられる．他にも，“～より…のほうが ” といったような評価対象，比較対象が確実に同定できるような構文パターンを用いて，比較対象集合を取得するという手法が考えられる．これらの手法によって自動取得する際には，ノイズが混じらないようにうまく処理する必要がある．

6. ま　と　め

本稿では，Web からの比較文マイニング及び比較文を用いた補間エンティティの発見手法を提案した．提案手法では，与えられた 2 エンティティとそれらに対応する比較対象集合を組み合わせ，比較文に特有な構文パターンを用いたクエリを生成し，Web からの比較文の抽出を行う．そして，得られた比較文から

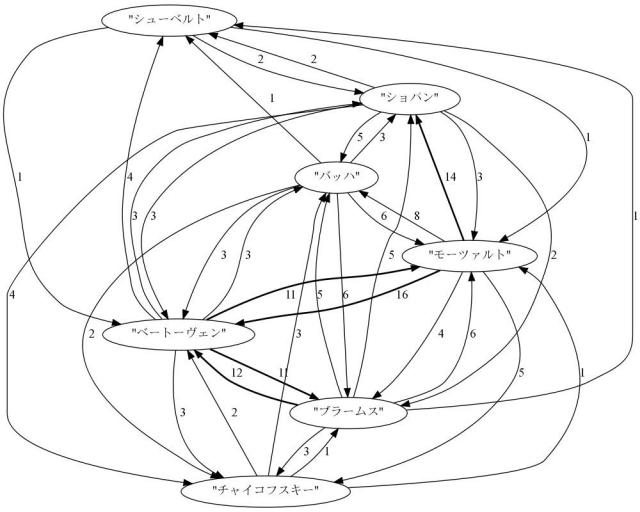

図 9　音楽家に関して作成したグラフ

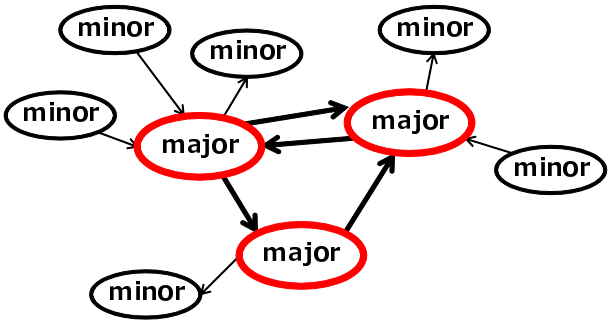

図 10　メジャー・マイナー関係が存在すると起こりうる現象

用意した辞書とのマッチングを行うことで，評価対象，比較対象，評価の極性を獲得し，グラフを生成する．そしてグラフから補間エンティティ及びその系列の発見を行う．我々はグラフから補間エンティティ及びその系列を発見する 3 つの手法を提案し，実験によりそれぞれの有効性，特徴を確かめた．また，好みやおいしさといった観点における補間エンティティ及び系列を比較文によって抽出できることが示せた．しかし，比較文で表されないようなマイナーなエンティティを含んだ集合に対する補間エンティティの発見は困難であることも明らかになった．今後，そういったエンティティ集合にも対応できるような手法を考案していく予定である．また，これまで 2 つのエンティティの間を発見するということをターゲットに取り組んできたが，このことはある観点における順序を考慮したエンティティ検索で有用である．補間エンティティを求めたいような時は，1 章で挙げたような例の場合や，求めたいエンティティをさらに絞り込む場合である．今後の展開として，1 つのエンティティを与え，そのエンティティよりある観点において優れている，あるいは劣っているエンティティを提示する検索，つまりへと絞りこむ前の段階に取り組んでいくことを考えている．

謝辞 本研究の一部は，グローバル COE 拠点形成プログラム “知識循環社会のための情報学教育研究拠点”，計画研究 “情報爆発時代に対応するコンテンツ融合と操作環境融合に関する研究”(研究代表者：田中克己，A01-00- 02, 課題番号 18049041), 未踏 IT 人材発掘・育成事業 2009 年度上期未踏ユースによるものです．ここに記して謝意を表すものとします．

文　　献

[1] Naoto Asahi, Takehiro Yamamoto, Satoshi Nakamura, and Katsumi Tanaka, “Finding Intermediate Entity between Two Examples on the Web”, In Proc. of the 11th ACM International Workshop on Web Information and Data Management (WIDM2009), pp.83-86, 2009.

[2] 旭 直人, 山本 岳洋，中村 聡史，田中 克己, “挟みこむ検索：明示的に与えられた観点に基づく補間エンティティの発見”, 情報処理学会創立 50 周年記念全国大会, 5ZN-5, 2010.

[3] Z. Ghahramani, and K. Heller, “Bayesian sets”, In Proc. of the 19th Annual Conference on Neural Information Processing Systems , vol.18, pp.435, 2006.

[4] 山口 雅史, 大島 裕明, 小山 聡, 田中 克己, “サーチエンジンのクエリログを利用した同位語の発見”, DBSJ Letters, Vol.5, No.2.

[5] D. Lin, “Automatic Retrieval and Clustering of Similar Words”, In Proc. of the 36th annual meeting on Association for Computational Linguistics, , pp. 768-774, 1998.

[6] K. Shinzato, K. Torisawa, “A Simple WWW-based Method for Semantic Word Class Acquisition”, In Proc. of the Recent Advances in Natural Language Processing (RANLP05), , pp.493-500, 2005.

[7] H. Ohshima , S. Oyama, and K. Tanaka, “Searching Coordinate Terms with Their Context from the Web”, In Proc. of the 7th International Conference on Web Information Systems Engineering, pp. 40-47, 2006.

[8] Google Sets http://labs.google.com/sets

[9] R. C. Wang, and W. W. Cohen, “Language- Independent Set Expansion of Named Entities Using the Web”, In Proc. of The IEEE International Conference on Data Mining , pp342-350, 2007.

[10] M.J. Cafarella, D. Downey, S. Soderland, and O. Etzioni, “KnowItNow: fast, scalable information extraction from the web”, In Proc. of the conference on Human Language Technology and Empirical Methods in Natural Language Processing, pp.563–570, 2005.

[11] O. Etzioni, M.Cafarella, D.Downey, A-M. Popescu, T. Shaked, S. Soderland, D.S. Weld, and A. Yates, “Unsupervised Named-Entity Extraction from the Web: An Experimental Study”, Artificial Intelligence, vol.165, pp.91- 134, 2005.

[12] P.P. Talukdar, T. Brants, M. Liberman, and F. Pereira, “A context pattern induction method for named entity extraction”, In Proc. of the Tenth Conference on Computational Natural Language Learning, pp.141–148, 2006.

[13] B. Liu, M. Hu, and J. Cheng, “Opinion observer: Analyzing and comparing opinions on the web”, In Proc. of the 14th international conference on World Wide Web, pp.342–351, 2005.

[14] N. Jindal, and B. Liu,“ Identifying Comparative Sentences in Text Documents”, In Proc. of the 29th annual international ACM SIGIR conference on Research and development in information retrieval, pp.244–251, 2006.

[15] N. Jindal, and B. Liu, “Mining comparative sentences and relations”, In Proc. of the National Conference on Artificial Intelligence , vol.21, pp1331, 2006.

[16] T. Kurashima, K. Bessho, H. Toda, T. Uchiyama, and R. Kataoka, “Ranking Entities Using Comparative Relations”, Database and Expert Systems Applications, pp.124–133, 2008.

[17] 佐藤敏紀, 奥村学, “blog からの比較関係抽出”, 情報処理学会研究報告, 自然言語処理研究会報告, No.94, pp.7–14, 2007.

[18] 鈴木智美，“比較表現 「A は B より～」 再考– 日本語教育における的確な導入例を考える–”，第 7 回日本語教育研究集会予稿集，pp.38–41，2009.