比較評価情報の抽出とそれに基づくランキング手法の提案

倉島　健† 別所　克人† 内山　俊郎† 片岡　良治†

† 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
〒 239-0847 神奈川県横須賀市光の丘 1-1
E-mail: †{kurashima.takeshi,bessho.katsuji,uchiyama.toshio,kataoka.ryoji}@lab.ntt.co.jp

あらまし　比較評価文は対象間の優劣を述べた表現である．対象（典型的には商品）を体験した消費者が発信する大量のテキストデータからこれらの情報を収集し分析することができれば，本当に価値のある対象を効果的に発見することができるはずである．本稿では，Consumer Generated Media(CGM) から人々が対象間の優劣を述べた比較評価情報を抽出し，それをもとに対象をランキングする手法を提案する．対象をグラフのノードで，比較評価要素の抽出結果集合から導き出された対象間の優劣関係をグラフの有向辺で表現したグラフを生成し，その構造を解析することで各々の対象の評価値を算出する．このグラフは「より良い商品を求めて売り場を移動する顧客」の行動をモデリングしたものである．

キーワード　データマイニング，Web とインターネット，知識発見

Ranking Method using Comparative Relations extracted from CGM

Takeshi KURASHIMA†, Katsuji BESSHO†, Toshio UCHIYAMA†, and Ryoji KATAOKA†

† NTT Cyber Solutions Laboratories, NTT Corporation
Yosidahonmati,Hikari-no-oka, Yokosuka-shi, Kanagawa, 239-0847 Japan
E-mail: †{kurashima.takeshi,bessho.katsuji,uchiyama.toshio,kataoka.ryoji}@lab.ntt.co.jp

Abstract　A comparative sentence expresses a relation between two objects. By mining the information from a large number of documents generated by consumers, we can rate the values of objects efficiently. This paper proposes a method of mining the comparative relations and ranking the objects based on mined results. We generate the graph, each node corresponds to a object, and each edge expresses the relation between two objects. By applying the PageRank method to the graph, the system brings order to the object. This type of graph can be thought of as modeling the behavior of a potential customer.

Key words　Data Mining, Web and Internet, Knowledge Discovery

1. はじめに

掲示板や Blog には，人々が商品等の対象を実際に使用し述べた感想が高い頻度で記述されている．従来，消費者は商品を提供する側から一方的に発信される広告情報に頼ってきたが，商品を実際に体験した人々の反応，いわゆる ”口コミ ”情報を参照して商品を選択するようになってきている．ある商品を体験した個人の情報発信は，商業的な意図で書かれたコンテンツとは別の形で有意義である．

個人が発信するこのような文書の中には，個人が体験した複数の商品の優劣を述べた比較評価文が存在する．例えば，「A のデザインは B よりも良い」といった表現である．この文はある対象の価値を他の対象との相対的な位置づけによって表現したものであり，複数の対象の関係性を記述しているという面で非常に価値のある情報といえる．消費者が発信した大量のテキストデータからこれらの情報を収集しまとめあげることができれば，本当に価値のある対象（商品）を発見できるはずである．

現在，Web 上に分散している様々な文書から意見を記述している箇所を抽出し，それらを要約してレーダーチャートのような形式で可視化するシステムが存在する．これらのシステムは，一般に対象，属性，評価という 3 つ組の意見モデルを定義して情報抽出を行っている．つまり，個々の対象の単体評価を求めているため，複数の対象の比較検討は，個々の対象に関して得られた意見の分布を参照するにとどまっていた．

本研究においては，人びとが複数の対象を比べあわせて，そこに認められる優劣を述べた比較評価情報を抽出する．比較評価情報は {評価対象，比較対象，属性，評価} から構成され，本研究においては，このうちの {評価対象，比較対象，評価}

に焦点を絞って抽出を行う．属性の抽出については，従来の３つ組の意見抽出における主要課題であり，今後はこれらの研究を参考にして抽出する予定である．本研究では，さらに，これによって得られた２つの対象間の関係性から対象空間全体におけるそれぞれの対象の価値を算出する．対象をグラフのノードで，比較評価要素の抽出結果から導き出された対象間の関係をグラフの有向辺で表現したグラフを生成し，その構造を解析することで各々の対象の評価値を算出する．より直感的には，このグラフは「より良い商品を求めて売り場を移動する顧客」の行動をモデリングしたものである．得られた結果は，商品の購入を検討している潜在的な顧客や，マーケットアナリスト，あるいは商品を提供する側にとって有益な情報である．

以降，２章で関連研究について，３章で本手法の概要について，４章で比較評価要素抽出について，５章で相関ルール分析について，６章でランキング手法について，７章で評価実験の結果について，８章でまとめと今後の課題について述べる．

2. 関 連 研 究

評判に関する研究においては，文書分類と情報抽出という二つの大きな流れが存在し，本研究は，後者の情報抽出分野に属する．従来の情報抽出分野における主要な課題は，評価情報を構成する３つ組（対象，属性，評価）の抽出である．立石らは対象，属性，評価に関する共起パターンを介して，評価表現と属性表現をブートストラップ的に抽出する手法を提案した [1] [2]．この手法によって作成した評価表現辞書，属性表現辞書と手動で作成した抽出ルールとをマッチングすることで３つ組の抽出を行っている．Bing [3] らは，３つ組における属性の抽出をテキストマイニング分野における相関ルールマイニングを利用して抽出している．また，いずれの研究においても，抽出した３つ組を可視化し提示するインタフェースを持つシステムを提案している．これらの研究は，単一の対象に対する評価を扱っているといえる．一方で，複数の対象を相対的に評価した比較評価文に着目した情報抽出も行われ始めている．Nitin [4] [5] らは，相関ルールを拡張した手法により，比較評価要素を抽出するパターンを自動生成する手法を提案している．本手法においても，Nitin らと同様，比較評価要素の抽出を試みているが，人手で作成したパターンと対象候補集合との照合を組み合わせることで，Nitin らの手法と同程度の精度で比較評価情報における対象と比較対象の抽出を実現した．また，本研究は，比較評価要素を抽出するに留まらず，抽出結果集合から対象間関係のマイニングを試みている．

3. 比較評価マイニング

複数の対象を正しく順序付けることができるのはその両者を体験した人のみである．ある対象に対する単体評価は，それ自体の価値を推し量るにはある程度有用だが，対象集合の中におけるその対象の価値を量るのに十分な情報ではない．本研究の目的は，大量のテキストデータから比較評価要素を抽出し，それをもとに対象集合の中から本当に価値のあるものを発見することである．この一連の流れを「比較評価マイニング」と呼ぶこととする．本手法は以下の３ステップから構成される．

- 比較評価要素の抽出
- 相関ルール分析
- ランキング

比較評価マイニングの最初のステップはテキストデータからの比較評価要素の抽出である．比較評価要素をパターンマッチングと対象集合との照合を組み合わせた手法で抽出する．次に，比較評価要素を格納したデータベースからの相関ルール分析によって２つの対象間の優劣関係を導き出し，それをもとに対象空間全体を表現するグラフを生成する．最後に，生成したグラフの構造を解析し，対象空間全体の中における，ある対象の価値を求める．以降，それぞれのステップについて述べる．

4. 比較評価要素の抽出

本章では，ユーザが比較したい対象集合（典型的には競合商品など）を入力として，それに関連する比較評価要素をテキストから抽出する手法について述べる．以降，比較評価要素の定義とその抽出手法について述べる．

4.1 比較評価要素の定義

比較文とは，複数の対象の相対的な関係性について述べた文である．Nitin らの定義によると，英語における比較文は以下の４つのタイプに分類できる．

- Non-Equal Gradable:ある属性に関して複数の対象を優劣付けた表現（例）greater，less than，-er -est.
- Equative:ある属性に関して２つの対象が等しいことを述べた表現（例）equal to，as as.
- Superlative:複数の対象の中である対象がもっとも優れていることを述べた表現（例）Greater or less than all others.
- Non-Gradable:ある属性に関して複数の対象を比較した表現だが，明示的に優劣付けていない表現（例）Toyota has GPS, but Nissan does not have.

本研究においては比較文における４つのタイプの中で特に Non-Equal Gradable に着目して抽出を行う．本稿では，Non-Equal Gradable に対応する日本語の表現を比較評価文，比較評価文を構成する要素を比較評価要素と呼ぶことにする．比較評価要素を構成する要素を以下に示す．

{評価対象，比較対象，属性，評価}

比較されている二つの要素の中で評価表現に意味的に主格で係っている方を「評価対象」，係っていないほうを「比較対象」とする．また，評価対象と比較対象はユーザから入力として与えられる対象集合と同じ概念階層に属する語でなければならない.「属性」は，評価対象と比較対象を比較する際の評価項目であり，両者に共通する性質や特徴を示す表現である.「評価」は２つの対象間の優劣を述べた表現である．また,「評価」に関連が深い要素として,「極性」がある.「極性」は「評価」が肯定か否定のどちらの意味を持つかを示すものであるが，テキストから抽出を行う情報ではないので, ここでは比較評価要素に含めないこととした.

英語においては，ほとんどの形容詞・副詞に対して，比較級，最上級が存在する一方，日本語においては，比較級・最上級と

いうものは存在せず，英語のような形容詞・副詞の変化はない．その代わりに「より」や「ほうが」などの比較評価文に特有な表現を用いて示される場合が多い．典型的には，評価対象が「ほうが」などの表現とともに出現し，比較対象が「より」や「と比べ」などの表現とともに出現する．本手法においてはこれらの表現を手がかりに抽出を行う．なお，本研究においては，比較評価要素の中でも特に {評価対象，比較対象，評価} の抽出に焦点を絞って抽出を行う．

4.2 比較評価要素抽出アルゴリズム

本章では比較評価要素の抽出手法について述べる．基本的な抽出アルゴリズムは以下のステップからなる．

（1） ユーザは入力として初期対象集合（ルート集合）$S = \{s_1, ..., s_m\}$ を与える．典型的には，ルート集合はユーザが比較を行いたい複数の対象である．

（2） システムはルート集合に関連する Web 文書集合を収集する．文書を形態素解析し，解析データ $P = \{p_1, ..., p_n\}$ を生成する．

（3） P のそれぞれの要素 p_i から以下の形式のトランザクションを生成する．

$T = (x_1, .., x_m, y_1, .., y_n, t_{x_1}, .., t_{x_m}, t_{y_1}, .., t_{y_n}, \text{評価}, \text{極性})$

x: 評価対象の候補語
y: 比較対象の候補語
t: クラス値

クラス値は比較評価表現に特有のパターンで抽出された語とそうではない語を判別するためのものである．

（4） ルート集合 S からベース集合 S' への拡張を行う

（5） トランザクション T の評価対象 x_i，及び比較対象 y_i とベース集合 S' とを照合する

（6） システムは $T' = (x', y', \text{評価}, \text{極性})$ の集合を出力する

以降，それぞれの抽出ステップの詳細について述べる．

4.2.1 Web 文書の収集と解析データの生成

ユーザからの入力 S（ルート集合）に基づいて Web 文書を収集する．主要なポータルサイトにおいては，実際に商品を購入した顧客が書いたユーザレビューが大量に投稿されている．また，そのレビューが何の対象に関して書かれたかがメタ情報として与えられている場合が多い．これらの情報を利用して，初期対象集合に関連する記事をクローリングする．また，ブログを解析対象とする場合は，S の各要素をブログサーチエンジンのクエリとして投げ，検索結果からブログの本文情報を取得する．次に，収集したテキストから解析データ $P = \{p_1, ..., p_n\}$ を生成する．このデータはテキストを形態素解析した結果に基づいて作成するものである．以下にデータの形式を示す．

$p_i = [\text{単語}_1/\text{形態素}_1][\text{単語}_2/\text{形態素}_2]\cdots[\text{単語}_o/\text{形態素}_o]$

形態素解析結果をもとに，テキストを文分割する．P の各要素 p_i はひとつの文に相当し，比較評価要素の抽出は p ごとに行う．

4.2.2 評価表現の抽出と極性の付与

評価表現の抽出は抽出パターンと評価表現辞書との照合に基づいて行う．評価表現抽出パターンの一例を表 1 に示す．これらのパターンにマッチする文字列を評価表現として抽出し，パターンの下線部分にマッチする文字列に対して評価表現辞書照合を行う．辞書にその語が存在した場合にはこのパターンにマッチした文字列を評価表現と判定する．評価表現辞書に格納されているデータの形式は {単語，形態素情報，極性} である．辞書照合を行うと同時に，極性の値も取得する．極性は肯定か否定のどちらかの値をとる．また，この抽出パターンは「退屈しない」や「面白くない」といった打ち消しの表現「ない」も，評価表現の一部として抽出するという特徴がある．打ち消しを含む抽出パターンで評価表現が抽出された場合には，評価表現辞書の照合によって得られた極性の値を反転させる．

表 1　評価表現抽出パターンの一例

[*/動詞語幹][*/動詞活用語尾][*/動詞接尾辞:終止]
[*/形容詞語幹][*/形容詞接尾辞:終止]
[*/形容詞語幹][*/形容詞接尾辞:連用][*/連用助詞]? [な/形容詞語幹:存在]([*/形容詞接尾辞:終止]|[*/形容詞接尾辞:連用])
[*/名詞：形容]([の/連体助詞][よう/補助名詞])?[*/判定詞：終止]
[*/名詞：形容]([の/連体助詞][よう/補助名詞])?[*/判定詞：連用] [な/形容詞語幹:存在]([*/形容詞接尾辞:終止]|[*/形容詞接尾辞:連用])

表 2　評価対象・比較対象抽出ルールの一例

[*/名詞]([に/助詞]|[と/助詞])[比べ/動詞語幹]→ 比較対象,T
[*/名詞]([に/助詞]|[と/助詞])[比較/名詞：動作]→ 比較対象,T
[*/名詞][より/格助詞]→ 比較対象,T
[*/名詞][に/助詞]→ 比較対象,F
[*/名詞][の/連体助詞][ほう/補助名詞]([が/連用助詞]|[は/連用助詞]) → 評価対象,T
[*/名詞]([は/助詞]|[が/助詞])→ 評価対象,F

4.2.3 ルールマッチングによる評価対象・比較対象候補語の抽出

ルールマッチングにより，評価対象と比較対象の候補語を抽出する．評価対象抽出ルールと比較対象抽出ルールの一例を表 2 に示す．ルールの形式は以下のようなものである．

抽出パターン　→ { 要素，クラス }

ルールの条件部には抽出パターンが，結論部には要素名とクラス値の組を記述する．抽出パターンは評価表現抽出パターンと同じ形式であり，このパターンと解析データ p とを照合することによって抽出を行う．結論部の要素名は,「評価対象」か「比較対象」のどちらかの値をとる．クラス値は，次のステップ（対象集合の拡張）のために付与されるデータであり，このクラスが T の値をとるルールは，比較評価表現に特有な表現に基づくものである．例えば,「より」「比べ」や「ほうが」などの表現を用いた抽出パターンがこれに該当する．クラスが F の値をとるルールは，比較評価表現に特有な表現ではないが，評価対象，及び比較対象が出現する可能性があるパターンを示すものである．ルールの条件部の * にマッチする語が，結論部に示す要素の値として抽出されることになる．1 文内で複数の評価対象，あるいは比較対象が抽出される場合は，両者とも候補語として抽出する．ここまでの解析結果から以下のような形式のトランザクションを生成する．

$T = (x_1, .., x_m, y_1, .., y_n, t_{x_1}, .., t_{x_m}, t_{y_1}, .., t_{y_n}, \text{評価}, \text{極性})$

x: 評価対象の候補語
y: 比較対象の候補語
t: クラス値

例:「出演者はミッション・インポッシブル２のほうが１よりも＜評価＞良い＜/評価＞」という文を解析する場合を考える.「良い」は評価表現抽出のステップですでに抽出されており,「肯定」の極性が付与されているものとする.「１よりも」は，要素値が「比較対象」，クラス値が「T」を持つルールの抽出パターンにマッチする．よって「１」を比較対象の候補語として抽出する.「出演者は」と「ミッション・インポッシブル２のほうが」は両者とも，要素値が「評価対象」を持つルールの抽出パターンにマッチする．しかし，クラスの値が前者が「F」，後者が「T」であるという違いがある．これは後者が「ほうが」という比較評価表現に特有な表現に基づいた抽出パターンで抽出された語だからである．最終的にこの文から生成されるトランザクション T は以下のものである．

T=(出演者, ミッションインポッシブル 2,1,F,T,T, 良い, 肯定)

4.2.4 対象集合との照合

比較評価要素抽出の最後のステップとして，前節までに得られた評価対象，及び比較対象の候補語と対象集合に含まれる要素とを照合する．対象集合との照合を行うことで以下のような利点がある．

- 求めたい概念階層の比較評価要素のみを抽出可能
- 評価対象，及び比較対象抽出精度の向上

現段階のトランザクション集合の中には様々な概念階層の比較評価要素が含まれている．例えば，映画のユーザレビューの中には映画間の比較もあれば俳優間の比較もある．トランザクション集合の中から，ユーザからの入力である初期対象集合 S に関連する比較評価要素のみを抽出する必要がある．また，パターンのみでは評価対象を抽出することが困難な場合が存在する．以下の２つの例文においては，評価対象である「映画A」と属性である「俳優」がまったく同じパターンで出現している．このような場合でも，対象集合との照合を行うことで「映画A」のみを評価対象として正しく抽出することが可能となる．

- 例１　映画Aは映画Bよりも良い
- 例２　俳優は映画Bよりも良い

しかし，この手法では初期対象集合間の比較評価要素しか得ることができない．そこで，本手法においては，前節までに得られた評価対象，及び比較対象の候補語と対象集合との照合を行う前に，対象集合を拡張する．この拡張は，例えば「A より B のほうが 良い」というような評価対象，比較対象がともに比較評価文に特有な表現で出現した文章を対象に行われる．この場合，AとBは高い確率で同じ概念階層に属する語であり，どちらか一方の要素が対象を示す表現であれば，もう一方の単語も高い確率で対象を示す表現であるといえる．このようにして対象を拡張することにより，初期対象集合 S に含まれない語との比較評価情報も抽出することが可能となる．その関係を図 1 に示す．この拡張は後述するランキング手法において効果をもたらすものである．対象間の直接的な比較評価情報が得られない場合でも，そのほかの対象を介して両者の優劣関係が導かれる場合があるためである．以降，対象集合の拡張ステップについて述べる．

（1） 対象集合 S' に初期対象集合 S を追加
（2） トランザクション T を読み込む
（3） x_i と y_i がともに比較評価表現に特有な表現で抽出された語 (クラス値がT) であり，かつ，どちらか一方がすでに対象集合に存在する場合のみ，もう一方の要素を対象集合 S' に追加する．
（4） 2 へ戻る．読み込むトランザクションがなくなった場合は 5 へ．
（5） 2 から 4 のステップを数回繰り返す．

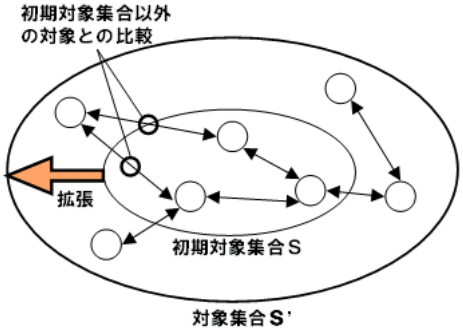

図 1　ルート集合 S とベース集合 S'

対象集合の拡張の後，前節までに得られた評価対象，及び比較対象の候補語と対象集合 S' とを照合する．以降，照合の各ステップについて述べる．

（1） トランザクション T を読み込む
（2） x_1 から x_m までのそれぞれの要素について，対象集合 S' に含まれる要素かどうかをチェックする．存在すればその語を最終的な評価対象 x' とする．
（3） 比較対象候補 y_1 から y_n についても同様．対象集合 S' に存在する語を最終的な比較対象 y' とする．
（4） (2) と (3) の結果に基づき，$T' = (x', y', \text{評価}, \text{極性})$ を生成.
（5） 1 へ戻る．読み込むトランザクションがなくなった場合は終了.

このようにして得られる，$T' = (x', y', \text{評価}, \text{極性})$ が最終的に得られる比較評価要素である．ただし，１文内から複数の評価対象，比較対象が抽出された場合は，　タイプがTの語　評価表現との距離が近い語，を優先的に抽出する．また，タイプTで抽出された評価対象，もしくは比較対象が存在しない要素に関しては比較評価文でない可能性が高いので削除する．

5. 相関ルール分析

本章では，次章の比較評価情報に基づくランキングの前処理として相関ルール分析技術を用いて，２対象間の優劣関係を導き出す．相関ルール分析は，ある商品 A を買ったら他の商品 B も同時に買うというような暗黙的なルールをマイニングする手法である．相関ルール分析では，アイテムの集合であるトランザクションを対象にして，各トランザクションの中に出現する

アイテム組に見られる共起パターンを発見する．相関ルールは以下のように表現できる．　　$X \Rightarrow Y$

X：条件部

Y：結論部

条件部と結論部は，複数のアイテムであってもよい．この相関ルールは，もし条件部のすべてのアイテムがトランザクション中に現れれば，そのトランザクション中には結論部のすべてのアイテムが現れるという意味となる．次に，支持度という概念について説明する．データベース D 中においてアイテム集合 X と Y をともに含むトランザクションの全トランザクションに対する割合を支持度（support value）といい，sup(X ⇒ Y) と表記する．一方，テキストマイニング分野においては，アイテム集合 X と Y をともに含むトランザクションの数を支持度と呼ぶことがある．本研究においてはテキストマイニング分野における定義を用いることとする．データベースに格納されている比較評価要素の形式をもう一度示す．

$T =$ (評価対象, 比較対象, 評価，極性)

極性: P(肯定) か N(否定) の 2 値

このデータベースから以下の値を算出する．

S_{ij}：対象$_i$ が 対象$_j$ よりも良いと述べた要素数 $(i \neq j)$

まず，データベースへの問い合わせによって対象集合 $O = o_1, o_2, ..., o_N$ を取得する．そして対角要素が 0 であるN次元正方行列 S を作成する．S_{ij} はこのデータベースからの相関ルール分析の結果から算出する．計算式を以下に示す．

$$S_{ij} = sup(O_i, C_j \Rightarrow P) + sup(O_j, C_i \Rightarrow N)$$

O_i は 対象$_i$ が評価対象要素として出現した場合を，C_i は 対象$_i$ が比較対象要素として出現した場合を示している．この計算式で作成した行列 S の i 行目の要素の和は 対象$_i$ が優れていると述べられた総数であり，i 列目の要素の和は 対象$_i$ が劣っていると述べられた総数となる．

6. 比較評価情報に基づくランキング手法

前章の相関ルール分析で得られた 2 つの対象に関する局所的な二項関係は，ユーザがそれらを比較検討したいというケースにおいて有益な情報といえる．本章では，さらに二項関係から対象空間全体におけるそれぞれの対象の価値を導き出す手法について述べる．本手法で導入する仮説は以下のものである．

- 多くの良質な対象と比較されて相対的に評価を得ている対象はやはり良質な対象である

この仮説における「相対的な評価」とは，2 つの対象の優劣の度合い・程度を表すものである．また，良質な対象と比べて相対的に評価されている対象は，やはり良質であり，比較される対象が多ければ多いほどその対象の価値は上がる．本手法においては，大域的な視点における，ある対象の価値を導きだすために，1 つのノードが 1 つの対象を，ノード間の有向辺が対象間の比較関係を表現するグラフを構築する．そして，ノードからノードへと次々に値を伝播させていき，最終的にそれぞれのノードに滞留する値に基づいて評価値を算出する．このグラフは「より良い商品を求めて売り場を移動する顧客」の行動をモデリングしたものである．

6.1 グラフの生成

1 つのノードは 1 つの対象を表現し，ノード間の有向辺は対象間の関係性を表現する．有向辺は重み付きであり，あるノードから他ノードへの遷移確率は，それら対象（ノード）間の相対的な評価・優劣によって決まる．このグラフの遷移確率行列 A を以下に示す．今，ノード V の総数を N とする．A_{ij} はノード V_i からノード V_j に対する遷移確率であり，i と j は 1 から N の値をとる．

$$A_{ij} = \begin{cases} \alpha \frac{S_{ji}}{S_{ji}+S_{ij}} & \text{if}(i \neq j) \\ \beta \frac{W(i)}{G(i)} & \text{if}(i = j) \\ 0 & \text{if}(i \neq j \text{ かつ } S_{ij} = 0 \text{ かつ } S_{ji} = 0) \end{cases}$$

S_{ij}: ノード V_j に対するノード V_i の支持数

$W(i)$: ノード V_i が勝利数

$G(i)$: ノード V_i の対戦数

α: 他ノードへの遷移確率に対する重み

β: 自己遷移確率に対する重み

数式内で使われている用語の意味は以下の通りである．

- 支持数:対象 A と対象 B 間における A の支持数とは，B よりも A のほうが良いと述べた要素の数である．
- 対戦数:対象 A の対戦数とは，対象 A と比較された対象の総数である．
- 勝利数:対象 A と対象 B 間において，A の支持数が B の支持数よりも多かった場合に，対象 A は対象 B に勝利すると呼ぶ．対象 A の勝利数とは，対象 A が他の商品に勝利した回数である．

S_{ij} の値は前章の相関ルール分析で得られた結果を用いる．$W(i)$ と $G(i)$ は，それぞれ S から求めることができる．

$$W(i) = w_{i1} + w_{i2} + \cdots + w_{ij} + \cdots + w_{iN} \quad (j \neq i)$$

$$G(i) = g_{i1} + g_{i2} + \cdots + g_{ij} + \cdots + g_{iN} \quad (j \neq i)$$

$$w_{ij} = \begin{cases} 1 & \text{if}(S_{ij} > S_{ji}) \\ 0 & \text{if}(S_{ij} \leqq S_{ji}) \end{cases}$$

$$g_{ij} = \begin{cases} 0 & \text{if}(S_{ij} = 0 \text{ かつ } S_{ji} = 0) \\ 1 & \text{else} \end{cases}$$

以上の式から行列Aを作成し，最後に，行列の各要素を，その要素を含む行の和 $\sum_{j=1}^{N} A_{ij}$ で割る．

$$A_{ij} = \frac{A_{ij}}{\sum_{j=1}^{N} A_{ij}}$$

そうして得られた A が遷移確率行列であり，A_{ij} はノード V_i からノード V_j への遷移確率をあらわす．ノード V_i に対するノード V_j の支持数 S_{ji} が，ノード V_j に対するノード V_i の支持数 S_{ij} よりも相対的に多い場合に，$A_{ij} > A_{ji}$ となる．対戦数と勝利数とに基づいて算出する自己遷移確率 A_{ii} は，部分グラフが複数生成される状態になったときに，それぞれのクラスタにおいてもっとも評価の高いノードに重みを持たせるための

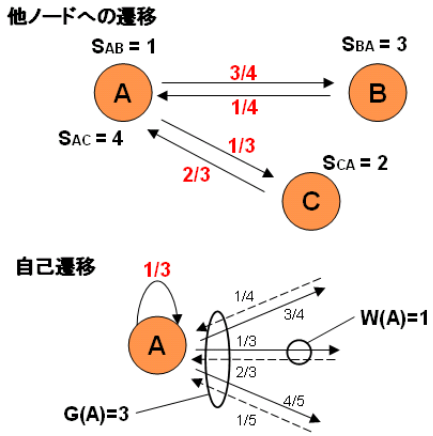

図 2　遷移確率の計算方法

ものである．図 2 に，シンプルな遷移確率の求め方を示す．

例: 比較評価要素集合から相関ルール分析によって以下のような S の値が得られたとする．

S_{AB}=1，S_{BA}=3，S_{AC}=4，S_{CA}=2, S_{AD}=1, S_{DA}=2

S_{AB}=1，S_{BA}=3 は対象 B よりも対象 A のほうが良いと述べている要素数が 1，その逆は 3 であるということを意味する．これらの結果を用いて行列を作成すると以下のようになる．ここでは $\alpha = 1$，$\beta = 1$ として計算する．例えば，A_{AB}=S_{BA}/(S_{AB}+S_{BA})=3/(1+3)=3/4 となる．

$$\begin{pmatrix} 1/3 & 3/4 & 1/3 & 2/3 \\ 1/4 & 1 & 0 & 0 \\ 2/3 & 0 & 0 & 0 \\ 1/3 & 0 & 0 & 1 \end{pmatrix}$$

そして，各ノードからの遷移確率の和が 1 となるように，各要素を，各行ベクトルの和で割る．

$$\begin{pmatrix} 4/25 & 9/25 & 4/25 & 8/25 \\ 1/5 & 4/5 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1/4 & 0 & 0 & 3/4 \end{pmatrix}$$

これが遷移確率行列Aである．この遷移確率行列を持つグラフを図 3 に示す．ノード A から出ている 3 つの有向辺について考えると，ノード A と比べて最も相対的に優れているノード B への重みが一番高くなる (9/25)．逆に，ノード A と比べて最も相対的に劣っているノード C への重みが一番低くなる (4/25)．

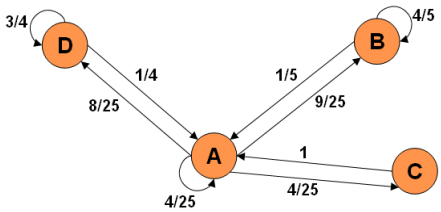

図 3　例題の遷移確率行列が表現するグラフ

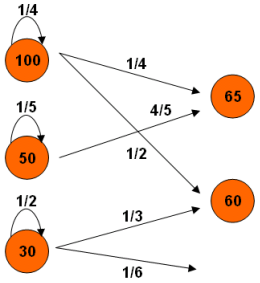

図 4　PageRank の算出方法

6.2　評価値の算出

生成したグラフ（遷移確率行列）から，それぞれのノードの評価値 $S(V)$ を求める．本手法においてはこの計算に PageRank を用いた．PageRank は本来，Web ページの重要度を求めるアルゴリズムであり，その計算法はグラフ理論に基づいている．Web ページをノード，Web ページ間のハイパーリンクを有効辺とし Web 空間をグラフに投影する．そして,「多くの良質なページからリンクされているページは，やはり良質なページである」という考えのもとにそれぞれのページの評価値を再帰的に求めるのが PageRank である．PageRank における評価式を以下に示す．

$$S(V_j) = (1-d) \times \Sigma(A_{ij} \times S(V_i)) + d$$

ノード V_j の評価値 $S(V_j)$ はそれにリンクを張っているノードの評価値から計算する．d は，ユーザが現在見ている商品（ノード）からまったく無関係な商品（ノード）の前に移動（ランダムジャンプ）してしまう確率である．PageRank をより直感的に述べると，ユーザがランダムに Web ページのリンクをたどり続ける行動のモデル（”random surfer”モデル）である．Web ページの評価値はランダムに Web 空間を歩き回り，最終的にそのページにたどりつく確率に等しい．

6.3　潜在的な顧客の行動モデル

生成したグラフをより直感的に述べると「より良い商品を求めて売り場を移動する顧客」の行動をモデル化したものといえる．顧客は商品群の中からある商品を見始め，次々に他の商品も見ていく．顧客の次の選択は大きく分けて次のふたつである．

（1）　現在見ている商品 A の前に留まり商品を見続ける

（2）　他の商品 B の前に移動する

この時，顧客が次にどのように行動するかは，今見ている商品と他の商品との優劣関係に大きく依存する．現在見ている商品 A よりも優れている商品が相対的に少ない場合には，顧客はその場所にとどまって今見ている商品を見続ける．逆にいうと，他に優れている商品が多く存在している場合には，他の商品の前に移動しやすい．

次に，2 の「他の商品 B の前に移動する」場合に，顧客がどの商品を選択するかについて述べる．顧客は現在見ている商品 A よりも相対的に支持されている商品ほど次に見る商品として選択しやすい．つまり,「$A < B$」と述べた人が「$A > B$」と述べた人よりも相対的に多い B ほど，顧客は次に見る商品として選択する．ある商品（ノード）に関して算出した評価値は，

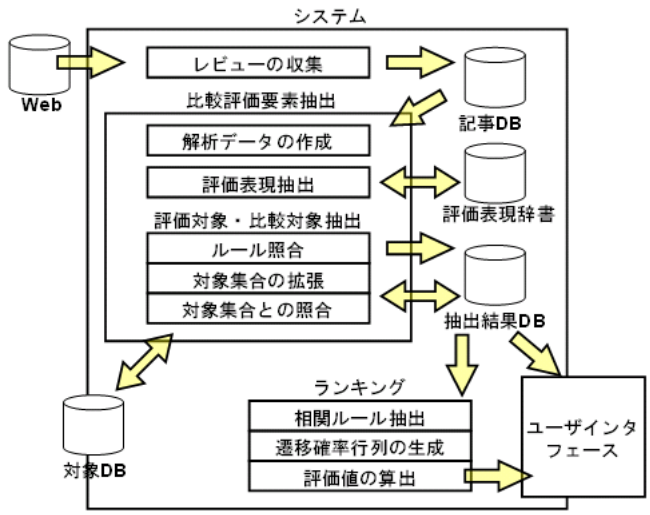

図 5　システムの構成

より良い商品（ノード）を求めて売り場をまわる顧客が最終的にその商品（ノード）にたどり着く確率に等しくなる．

7. 評　　価

7.1 システムの実装

提案した手法に基づき，プロトタイプシステムを実装した．システムの構成を図 5 に示す．文章の形態素解析には Jtag [9] を用いた．Jtag は文章を語に区切り，その語の形態素情報を出力する．また，評価表現辞書は人手で作成したものを用いている．本システムにおけるデータベースはすべて MySQL [6] を使用している

7.2 比較評価情報の抽出精度

実装したプロトタイプシステムに基づき，収集したデータから実際に比較評価要素の抽出を行い，評価対象と比較対象の抽出精度を評価した．具体的には，ルールマッチングと対象集合との照合を組み合わせることで，どの程度精度が改善されるかを観察した．以下に示す評価対象・比較対象抽出プロセスの組み合わせを考え，それぞれについて適合率，再現率，F-measure を算出した．

（1） ルールマッチング
（2） 映画辞書との照合
（3） 拡張した対象集合 S' との照合

映画辞書には 1995 年 1 月から 2006 年 12 月までに日本で上映された映画の正式名称が格納されている．対象集合 S' は，初期対象集合 S を拡張して得たものである．評価に用いたデータは以下の通りである．

- データ:映画のユーザレビュー
- 記事数:500（5 タイトル，各 100 記事)
- 初期対象集合 S:記事に関する映画 5 タイトル

最初にこれらの記事から人手で正解を作成する．その結果，500 記事中，63 記事に比較評価要素が含まれており，評価対象の要素数は 40，比較対象の要素数は 76 であった．その後，プロトタイプシステムを用いて抽出を行った．表 3 に抽出結果に基づいて計算した適合率，再現率，F-measure を示す．評価対象と比較対象ともに，ルールマッチングに加えて映画辞書と対象集合との照合を行った (1)+(2)+(3) の F-measure が最も高くなっている．これは，辞書や対象集合との照合を行うことでパターンのみの抽出に比べて適合率が大きく改善したことに起因するものである．また，(1)+(2) の映画辞書との照合においては再現率が 0.1 程度と思ったほど効果が出ていないことがわかる．一方で，対象集合との照合においては再現率が 0.5 程度である．これは決して映画間の比較が評価用データに少なかったのではなく，映画が正式名称で引用されることが少なく，省略された形や「前作」などの語で引用される頻度が非常に高かったためである．評価対象と比較対象の抽出精度を比較すると，評価対象抽出のほうがルールマッチングのみの場合に比べて精度が改善していることがわかる．これは，評価対象抽出においては特に，評価対象と属性がまったく同じパターンで出現することが多く，そういったケースで正解を抽出できているからである．

表 3　抽 出 結 果

		(1)	(1)+(2)	(1)+(3)	(1)+(2)+(3)
対象	適合率	0.377	0.333	0.833	0.714
	再現率	0.725	0.100	0.500	0.625
	F-measure	0.496	0.154	0.625	0.667
比較対象	適合率	0.737	0.889	1.000	0.981
	再現率	0.737	0.105	0.592	0.684
	F-measure	0.737	0.188	0.744	0.806

7.3 グラフの生成と映画のランキング

ブログから抽出した比較評価要素集合に基づき映画のランキングを行った．用いたデータは以下の通りである．

- 初期対象集合 S:期間中に上映中の映画 135 タイトル
- 拡張した対象集合 S':映画 462 タイトル
- 期間:2006 年 5 月 01 日〜2007 年 1 月 31 日

実験に用いたブログは，goo ブログから収集したデータである．それぞれのブログから比較評価要素の抽出を行い,「評価対象」，あるいは「比較対象」要素が欠けた場合には，ブログのタイトルに含まれる映画名を補足した．また，前章の評価において対象集合と映画辞書との照合を組み合わせた手法が最も精度が高くなったため，今回もその手法を用いた．対象集合の拡張によって得られた,「1」「2」「前作」や「原作」などの表現に関しては，現時点では人手で映画名あるいは「映画名の原作」というような変換を行っている．抽出した比較評価要素集合をもとに映画をランキングした結果を表 4 に示す．また，生成したグラフを図 6 に示す．このグラフは山田らの手法 [10] に基づいて作成した．グラフのノードはそれぞれ映画に対応しており，ノードをつなぐ有向辺はノード間の比較関係を示す．また，グラフの z 軸はそれぞれのノードの評価値に対応している．

7.4 考　　察

表 5 に，1 位となった映画「嫌われ松子の一生」に関して得られた結果を集計したデータの一部を示す．S_{AB} は，対象 A と対象 B 間における対象 A の支持数である．上位の語（下妻物語，原作）に対しては，相対的な優位さを示す値

表 4　映画のランキング結果 (d=0.3 α=1.0 β=1.0)

映画名	評価値
嫌われ松子の一生	0.0170
DEATH NOTE デスノート 前編	0.0141
ゲド戦記	0.0133
LIMIT OF LOVE 海猿	0.0132
武士の一分 (いちぶん)	0.0121
パイレーツ・オブ・カリビアン/デッドマンズ・チェスト	0.0120
硫黄島からの手紙	0.0119
カーズ	0.0100
ワイルド・スピード X3 TOKYO DRIFT	0.0100
どろろ	0.0094

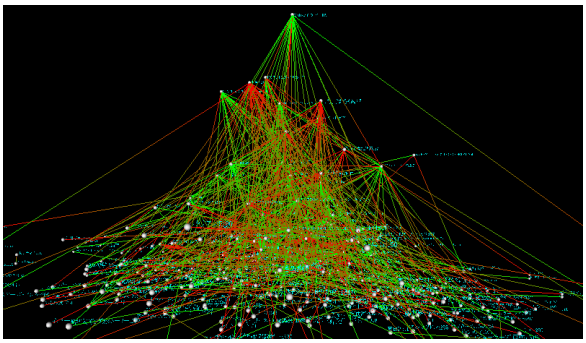

図 6　映画集合を表現したグラフ

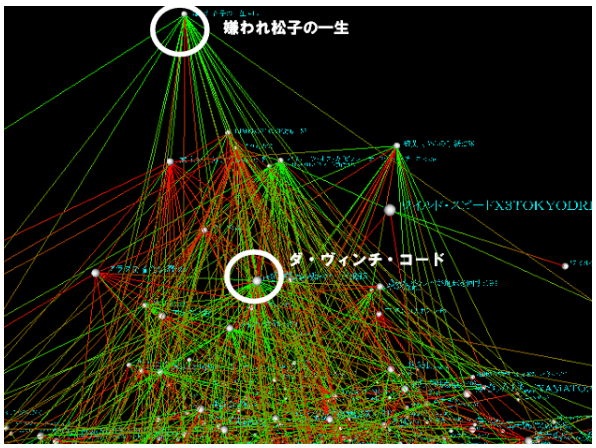

図 7　映画「嫌われ松子の一生」の周辺のノード群

(=$S_{AB}/(S_{AB}+S_{BA})$) はそれほど高くないが，23 の映画と比較され，そのうち 17 の映画に対して勝利（0.5 以上の値）しており，その平均値は 0.759 である．さらには,「ダ・ヴィンチ・コード」（評価値:0.0088）や「海猿ウミザル」（評価値:0.0057）といった評価値の高い映画に対して相対的に評価されており，それらのノードから伝播することで，高い評価値を得たといえる (図 7)．表 5 は，映画「ゲド戦記」に関して得られた結果である．比較関係は「アニメ」など，ジャンルが類似している映画間で得られる場合が多いことがわかる．また，上映期間が同じ映画間，監督が同じ映画間での比較評価が多く抽出された．可視化を行うと，ある側面に関して共通点（評価項目）を持つ映画の集合が形成されていることが確認できた．

表 5　映画「嫌われ松子の一生」に関して得られた比較評価情報

B	S_{AB}	S_{BA}
下妻物語	25	40
原作	21	29
ダ・ヴィンチ・コード	9	2
ドラマ	1	2
海猿ウミザル	2	0

表 6　映画「ゲド戦記」に関して得られた比較評価情報

B	S_{AB}	S_{BA}
ハウルの動く城	136	45
ブレイブストーリー	11	41
時をかける少女	4	39
原作	4	24
千と千尋の神隠し	12	9

8. ま と め

本稿では，ＣＧＭから比較評価要素の抽出を行った．さらに，対象をグラフのノードで，比較評価要素の抽出結果集合から導き出された対象間の優劣関係をグラフの有向辺で表現したグラフを生成し，その構造を解析することで各々の対象の評価値を算出する手法を示した．今後は属性（評価項目）の抽出も視野に入れてランキングを行っていく予定である．

9. 謝　　辞

日本電信電話株式会社 NTT コミュニケーション科学基礎研究所 斉藤和巳特別研究員には，提案手法の可視化に利用したグラフ構造可視化ツールを提供頂き，感謝します．

文　　献

[1] 立石健二，福島俊一，小林のぞみ，高橋哲朗，藤田篤，乾健太郎，松本裕治．”Web 文書集合からの意見情報抽出と着眼点に基づく要約生成”，情報処理学会自然言語処理研究会 (NL-163-1), pp.1-9.
[2] 立石健二，石黒義英，福島俊一．”インターネットからの評判情報検索”，情報処理学会自然言語処理研究会 (NL-144-11).
[3] Bing Liu, Minqing Hu and Junsheng Cheng. ”Opinion Observer: Analyzing and Comparing Opinions on the Web” In Proceedings of the 14th international World Wide Web conference (WWW-2005), May 10-14, 2005.
[4] Nitin Jindal and Bing Liu. ”Identifying Comparative Sentences in Text Documents” In Proceedings of the 29th Annual International ACM SIGIR Conference on Research & Development on Information Retrieval (SIGIR-06), 2006.
[5] Nitin Jindal and Bing Liu. ”Mining Comprative Sentences and Relations.” In Proceedings of 21st National Conference on Artificial Intellgience (AAAI-2006), July 16-20, 2006.
[6] MySql, http://www.mysql.com/
[7] L.Page, S.Brin, R.Motwani, T.Winograd. ”The PageRank Citation Ranking: Bringing Order to the Web”.
[8] S.Brin, L.Page. ”The Anatomy of a Large-Scale Hypertextual Web Search Engine”, http://www-db.stanford.edu/ backrub/google.html
[9] T. Fuchi, and S. Takagi, ”Japanese Morphological Analyzer using Word Co-occurence-JTAG”, COLING-ACL, pp.409-413, 1998.
[10] T. Yamada, K. Saito, and N. Ueda, ”Cross-Entropy Directed Embedding of Network Data”, Proc. of ICML'03, pp.832-839, 2003.