DEIM Forum 2014 F7-2

社会的印象に対する根拠の発見

海津　研† 加藤　誠† 大島　裕明† 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
E-mail: †{kaizu,kato,ohshima,tanaka}@dl.kuis.kyoto-u.ac.jp

あらまし　本稿では，人々が共通して抱く社会的印象に対して，なぜその印象を抱いたのか，という根拠になる情報を Web から抽出する手法を提案する．本稿では印象として扱える根拠を対象に，根拠として扱うべき‘対象’を社会心理学の知見を基に収集する．また，印象間の連想関係に着目し，人が印象を抱く過程のモデル化を行った．モデル化の観点から得られた根拠を，印象をどの程度連想させるのか，他の根拠をどの程度連想させるのか，といった評価指標から根拠度の定式化を行った．提案した根拠度を用いる事で，印象を抱く際により多くの人が共通して連想元として利用したと考えられる根拠が取得出来る．

キーワード　印象，根拠，関連語抽出，Web マイニング，PageRank

1. はじめに

人々は概して猫に対し “マイペース”，原子力発電所に対し “危険” といった印象を抱く．より多くの人が同様に考える印象を社会的印象と定義した上で，社会的印象は対象オブジェクトに抱かれた‘感じ’として捉える事が出来る．これは印象が「人間の心に対象が与える直接的な感じ」と定義がなされているためである (注1)．このような社会的印象は，元来オブジェクトに備わっていたわけではなく，オブジェクトに関する情報や，そのオブジェクトに関わる他の事象や印象を介して得られる．その際，どのような情報を基にそのような社会的印象を抱いたのかを知りたい，といった場面がしばしば見受けられる．例えばそのような状況として以下のような場面がある．

- どこかの米は “危ない”，という印象を抱いてしまった原因を探って，そのようなマイナスイメージを払拭したい．
- なぜ渋谷は “若者の街” と呼ばれているのか知りたい．
- 私の意見と異なり，世間では Ruby on Rails が簡単と思われているのはなぜか知りたい．

現状の検索エンジンでは社会的印象に対する根拠を得ることは難しい．これは，印象が他の根拠から連想されるためである．例として，渋谷が “若者の街” といった印象に対する根拠を取得する際，根拠には「東京にある」「ギャルが多い」「ファッション文化がある」，などが挙げられる．「東京にある」のは明確な事実である一方，「ギャルが多い」「ファッション文化がある」といった根拠については，事実であるかは人の主観に依る所が多く，事実であるかの判断はつかない．ここから根拠は印象であるかのように扱うことが出来る．検索エンジンはクエリを必要とし，検索結果は入力として与えられたクエリに関連する文書となる．しかし，先に述べたとおり根拠は印象であることも多く，根拠間にも連想関係が成立し，クエリの関連語を抽出するだけでは根拠の連想元となる根拠を発見する事が難しい．一方で，新しい語に関する知識が無い人々にとっては，直接オブジェクト名と印象をクエリに根拠となるような情報を返す検索エンジンが求められる．

根拠らしい根拠とは，より多くの人が社会的印象を連想する際に用いた根拠の事を指す．社会的印象がより多くの人が同様に考える印象であるのと同様に，より多くの人が印象を抱く際に参照するような根拠が，最も根拠らしい根拠であると考える．この時，根拠を印象として扱うとすると，根拠が印象を連想する以外にも，根拠が根拠を連想するといった事も考慮しなくてはならない．これは，原子力発電所が “危険” という印象における根拠に「人がガンになる」「放射性物質を拡散する」といったものがある際に，後者の根拠から前者の根拠が連想出来る事に対応する．

本研究では，初めに社会的な印象の根拠となりえる語を社会心理学の知見に基づいた上で収集し，根拠として扱うべき形で根拠の収集を行う．また人が印象を抱く過程をモデル化し，このモデルに基づき，根拠度の定義を行う．根拠度には，根拠がいかに印象を連想させるのか，また根拠がいかに他の重要な根拠を連想させるのか，といった点に着目する．

本稿の構成は以下の通りである．2 節では，印象及び関連語に関する関連研究を述べる．3 節では本研究における問題定義を行う．4 節では根拠集合の収集方法を説明し，5 節では，得られた根拠がいかに根拠らしいかを表す根拠度を求める手法を提案する．その語，6 節で提案した手法に関する実験及び評価を行い，最後に 7 節でまとめと今後の課題について述べる．

2. 関連研究

2.1 社会心理学における印象形成

印象に関する研究は，社会心理学の分野において古くから行われてきた．特に対人認知の分野において他者の印象がつくられるプロセスに着目した印象形成過程の理論化が行われきた．Asch はオブジェクトの印象は形容詞の評価を因子とした際に，全ての因子に対する寄せ集めでなく，複数の因子の布陣によって決定されるとするゲシュタルト心理学の立場から，印象形成理論を提言した [2]．また，理論化における別の考えとし

(注1)：goo 辞典 http://dictionary.goo.ne.jp/leaf/jn2/16434/m0u/

て，Anderson はオブジェクトの印象を因子に対する好ましさを統合したものとして考える事が出来るとした統合理論を提案した [1]．これらの研究は，対人認知処理の理論化に充分な役割を果たしている一方，どのような根拠が存在していたかの発見に関する研究としては扱われていない．

Brewer は，印象形成について対人認知処理モデルと呼ばれるモデルを用いて表している [9]．このモデルには認知を意識せずに行う自動的処理と，一定の判断が加わる統制処理が含まれている．この統制処理においてオブジェクトの印象に対してそのオブジェクトが属するカテゴリに当てはまるかどうかを考えるカテゴリ化と，一致しない場合に対する個別化といった処理が中で行われているといったモデルを Fiske は提言している [3]．これは新しいオブジェクトが与えられた際に，そのオブジェクトが属するカテゴリと似ているのか，似ていなければそのオブジェクト固有の印象だ，といった処理が行われている事に対応する．本稿では，この個別化とカテゴリ処理に基いて根拠の候補となるような語を取得する事を試みる．

2.2 関連語発見

根拠となる情報は，入力となるオブジェクトと印象を基に重要な関連情報を抽出する事に値するため，関連語の発見に関する研究が関連研究として挙げられる．TextRank [6] を用いると，得られたテキストから重み付きグラフを生成して，重要な関連語を発見する事が出来る．Turney らは 2 語間の関係性についての情報を教師無し学習で発見する手法を提案している [10]．稲川らは入力オブジェクトを特定するために必要な語集合の発見手法を提案している [12]．この手法では，候補語を取得した後に，入力オブジェクトに対する特定関係を語の絞込と主要度の 2 つの尺度を用いて評価している．

本研究では，根拠も印象であると考えているため，ある根拠が他の根拠を連想する連想関係が成立すると考えている．この連想関係について，良い根拠は他の良い根拠を連想させるもの，という再帰的な考えを用いて根拠度の指標に用いている．このような再帰性を用いて，複数ページのリンク構造に基づきページの重要性を計る PageRank [7] に関する研究は多数行われている．偏りをもたせた biased PageRank を用いて複数トピックに偏りを与えた topic-sensitive PageRank を Haveliwala [4] は提案している．本研究では，より印象を連想させる根拠が重要であると偏りを持たせている点で共通するものの，連想させるもの，つまりリンクを出す側のノードが重要であるといった点で異なっている．これは PageRank における重要度と捉えるより， Kleinberg [5] によって提案された HITS アルゴリズムの Hub が持つ特徴と似ているが， Authority に対応する値は無く，biased PageRank にある方向付きリンクに対して逆向きに重要度が流れる，といった表現が適切となる．

3. 根拠取得に関する問題定義

本節では，社会的印象の根拠に関する発見手法についての問題定義を行う．初めに社会的印象とは何か，社会的印象に関する根拠とは何かについての定義を行い，次に根拠集合の取得について説明し，根拠らしさを表す根拠度を計る問題を定義する．

3.1 社会的印象

印象とは「人間の心に対象が与える直接的な感じ」の事である．「原子力発電所が危険だ」といった文には ‘対象’ として原子力発電所が，‘感じ’ として “危険だ” といった値が対応付けされる．本研究では，印象を入力として受け付けるにあたり，‘対象’ であるオブジェクト $o \in T$ と，‘感じ’ である $i \in S$ の 2 つを入力に取るものとする．ただし，T は全ての ‘対象’ 集合，I は全ての ‘感じ’ 集合である．‘感じ’ とは，‘対象’ の状態や動作を表すものとする．

社会的印象として，本研究では万人受けされるような印象を対象とする．本来印象は個人が複数の ‘対象’ に対して複数持つものであるが，ある ‘対象’ に対して多くの人が同じ印象を持つことは良く見受けられる．このようなある ‘対象’ に多くの人が同様に考える印象についてを社会的印象として定義している．

3.2 社会的印象に関する根拠

まず初めに，本研究における根拠は，印象と等価なものとして扱う．これは印象に関する根拠が事実のみに依るものではないためである．根拠とは「物事が存在するための理由となるもの。存在の理由。」(注2) と定義がなされているため，物事の存在理由としての事実の因果関係が成立すれば良い事が分かる．しかし，印象が存在するための理由は事実のみならず，他の印象が基になる事が多い．これは渋谷が “若者の街” といった印象に対して「ギャルが多いから」といった根拠が存在している事からも分かる．この「ギャルが多い」という根拠は必ずしも万人に納得される事実にはならなく，‘対象’ としてギャル，‘感じ’ として “多い” を取る印象として表される．よって，本研究では根拠を印象と等価なものとして取り扱う．また，ここには印象間の連想関係が存在していると考えられる．これは，一方の印象を抱く事で他方の印象を抱くのは人間の心の中に連想が行われていると考えられるためである．本稿では入力として与えられた印象と根拠を区別をするため，入力として与えられ最終的に連想されるであろう印象についてのみを印象と呼び，それ以外の印象をすべて根拠と呼ぶ．

本研究において取得すべき社会的印象の根拠とは，社会的印象同様により多くの人が社会的印象を連想する際に用いた根拠の事を指す．このような根拠を人が見つける事で，結果としてより多くの人が印象を抱く事になる．これには，直接的に根拠から印象を連想するもののみならず，図 1 のように他の根拠を経由した上で印象を連想させる，といった事も考慮しなくてはならない．図 1 ではノードが根拠を表し，方向付きエッジにより連想を，太さが連想の強さを表している．

根拠は印象であることから，‘対象’，‘感じ’，及びその感じが肯定か否定かの 2 値，以上を含む 3 つ組みを用いて表現する．例えば，原子力発電所が危険といった入力が与えられた時，「放射性物質を拡散する」といった根拠に対しては，‘対象’ が放射性物質，‘感じ’ が拡散にあたり $(\text{放射性物質}, \text{拡散}, pos)$ のように表す．これは原子力発電所というオブジェクトに対し，放射性物質という ‘対象’ が何をするのか，どのような状態であるのかといった ‘感じ’ を人が抱いているのかを表現する．また文脈

(注2)：デジタル大辞泉 http://dictionary.goo.ne.jp/leaf/jn2/83302/m0u/

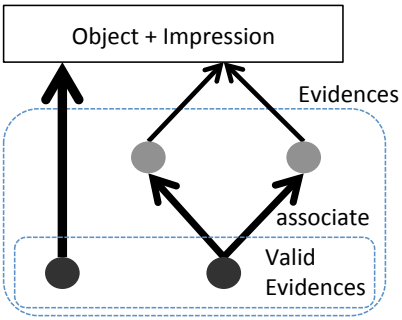

図 1　良い根拠から他の根拠を連想し，印象を連想する

上「放射性物質を拡散しない」といった文が出現する事もある．これは全く逆の ‘感じ’ を抱くのは明らかであるため，これらを区別すべく肯定か否定かを表す値も保持するものとする．本稿では根拠を $e \in E$ と定義し，$e = (t, s, b)$ のような 3 つ組みで表す．ここでは ‘対象’ を $t \in T$，‘感じ’ を $s \in S$，肯定文か否定文かの値を b と定義している．ただし，E を全ての根拠集合，$b \in \{\mathrm{pos}, \mathrm{neg}\}$ を取る．

3.3 問題定義

本小節で 3.1 小節で定義した印象に対し，3.2 小説で定義した根拠を取得する方法，及びいかに根拠らしいかを表す根拠度を求める問題の定義を行う．

3.3.1 根拠の収集

初めに，根拠を考える上でどのような ‘対象’ に着目すれば良いのかの抽出を図る．入力としてオブジェクト o と印象 i が与えられた際，o や i に関わる語を全て ‘対象’ として考えて良いのかといった問題が生じる．なぜそのような印象を持ったのか，といった質問に対し，全ての語に対して ‘感じ’ を抱くとは考えにくい．よって印象を抱くに至るには印象形成に関わる ‘対象’ のみに着目すれば良い．また，全ての語を ‘対象’ として取り扱うにはあまりにも現実離れし過ぎている．よって，本研究では，初めに入力 o 及び i から得られる根拠に対する ‘対象’ 集合 $T_{o,i} = \{t_1, t_2, \ldots\}$ を取得する事を行う $(T_{o,i} \subseteqq T)$．

次に，得られた ‘対象’ 集合を基に根拠集合を得る必要がある．これは，o の i という印象に対し，‘対象’ 毎にどのような ‘感じ’ が人によって持たれているのかを抽出すれば良い．各 ‘対象’ からは複数の感じが得られる．よって，$T_{o,i}$ を通して得られた根拠集合 $E_{o,i} = \{e_1, e_2, \ldots\}$ $(E_{o,i} \subseteqq E)$ を抽出する問題となる．ここで，$e_u (\in E_{o,i})$ に含まれる ‘対象’t_u は $t_u \in T_{o,i}$ を満たす．

3.3.2 根拠度

本稿では 3.2 小節で定義した根拠に基づき，根拠がいかに根拠らしいかを表す度合いを根拠度と定義する．3.3.1 項で得られた根拠 $e_i (\in E_{o_i})$ に対し根拠度 $\mathrm{Evid} : T \times I \times E \to \mathbb{R}^+$ を求め，o，i における根拠毎の根拠度を計る．以上から，本研究では次のような過程で根拠らしい根拠を発見する

（1） 入力としてオブジェクトと印象を受け付ける
（2） 根拠となる集合の収集
（3） 得られた根拠の根拠度を算出

4. 根拠の収集

本節では入力として得られたオブジェクトと印象から，根拠集合を収集する手法について説明を行う．初めに，根拠になりえる ‘対象’ についての特性について説明し，特性毎に ‘対象’ の収集を行う手法を説明する．最後に得られた ‘対象’ 集合から根拠集合を取得する手法についての説明を行う．

4.1 印象形成に関わる対象の特徴

本研究では根拠を収集するに辺り，次の 2 つの根拠を用いる．

- 総体的対象を持つ総体的根拠
- 個別的対象を持つ個別的根拠

総体的根拠は，オブジェクト固有の印象ではなく，オブジェクトが属するカテゴリにおいて注目される印象に基づく根拠である．原子力発電所が “危険” といった印象に対して「事故が多い」という根拠は，原子力発電所のみならず火力発電所や水力発電所においても “危険” という印象を抱く．これは社会心理学の印象形成において， カテゴリ化と呼ばれる処理に基づいたものである．カテゴリ化は Fiske [3] が提言した連続体モデルの中で，認知処理を軽減するために人が行っている行動であるとされる．個別的根拠は，カテゴリ化では捉えられない根拠を対象にするものであり，あるオブジェクト固有の知識から得られる根拠となる．原子力発電所が “危険” には,「放射線を拡散する」といった根拠がある．放射線は他の類似オブジェクトには本来関係の無い ‘対象’ であり，原子力発電所のみにおいて印象が持たれる根拠となっている．連続体モデルではこのような個別にのみ適用される処理を個別化と定義しており，本研究でもこの知見を基に根拠の収集を行う．

‘対象’ には言語的な特徴から名詞を用いる．根拠に含まれる ‘対象’ には何かしらの ‘感じ’ を抱くことになり，‘感じ’ が与えられる対象としては名詞が多いためである．文章の解析には MeCab(注3) を用いて形態素解析を行う．得られる名詞に関して，連続する複数の名詞は 1 語の名詞として扱う．これは連続した名詞により固有名詞が表現されている場合に，固有名詞に対する評価が一般名詞に対して評価されてしまう事を防ぐためである．また，一般的な語について SlothLib(注4) により公開されている日本語ストップワード辞書を用いて除外している．

4.2 対象の収集

前小節で述べた ‘対象’ の特徴に基づき，‘対象’ の収集を行う手法について説明する．初めにどのような情報源から ‘対象’ を取得すれば良いのかについての言及を行った上で，その情報源から ‘対象’ を取得する方法についての説明を行う．

4.2.1 対象取得における情報源

‘対象’ を収集するにあたり，情報源として様々なものが考えられる．入力として与えられたオブジェクト o と印象 i を基に，o をクエリとして Web 検索結果を取得し i を含む文書を対象として収集を行う，o や i 以外に “理由” といった語をクエリに加えて Web 検索結果を取得して収集を図るといったような事

(注3)：MeCab http://mecab.sourceforge.net/
(注4)：SlothLib http://www.dl.kuis.kyoto-u.ac.jp/slothlib/

も出来る．また，QA サイトには理由を示す文言が多く現れるのでは，といった仮定に基づき ‘対象’ を収集するといった事も考えられる．

本研究では，‘対象’ 収集の情報源として o と i をクエリに Web 検索結果用いる．このような手法を取る理由として，o のみをクエリとする場合，i とは全く関係の無い文書ばかりが得られてしまう事，o と i と “理由” といった語を AND 検索して探さない理由として，事実の因果関係に着目するのではなく印象の連想関係に着目をしたいため，語を絞込過ぎない必要性があるためである．また Web 検索を用いる理由としては Web 上にある全ての文書を対象として検索を図る事で，より多様な印象に関する意見を収集するためである．

4.2.2 総体的対象の取得

総体的対象は，印象形成の認知処理で行なわれるカテゴリ化に似た考えに基づいた，印象形成において着目すべき ‘対象’ の事である．総体的対象では原子力発電所が危険といった印象に対して「事故」，猫がマイペースといった印象に対して「性格」といったより抽象度の高い語を取得する．カテゴリ処理ではどのようなカテゴリに属しているのか，といった事が重要となるが，実世界においてカテゴリを考慮した比較が行なわれると考えるより，既に知っている類似オブジェクトをそのカテゴリにおける標準と捉える事が多いのではないかと考えられる．これは新たに 4 番打者として入団した無名な野球選手のカテゴリ処理では，一般的な 4 番打者の印象として類似した 4 番打者がどうであるか，といった印象を考慮した上でステレオタイプを形成するのではないのか，といった考えに基づいている．よって，総体的対象を取得する上で入力として与えられた o の類似オブジェクトを用いて，そのカテゴリにおいて印象が形成される要因となるような ‘対象’ の取得を図る．

類似オブジェクトの発見に関して，Web 検索のスニペットから言語パターンで発見する手法を用いる [13]．比較対象として用いられる類似オブジェクトはこのような並列助詞で接続される事が多いため，本研究が対象にする類似オブジェクトもこのような手法から取得出来ると考えられる．例として原子力発電所に並列助詞で接続される類似オブジェクトとして火力発電所が，渋谷と並列助詞で接続される類似オブジェクトには品川や新宿を発見する事が出来る．

入力オブジェクト o 及び o から得られた類似オブジェクト集合 C_o を用いて，総体的対象集合 $T^{aggr}_{o,i} = \{t^1_{aggr}, t^2_{aggr}, \ldots\}$ ($T^{aggr}_{o,i} \subseteqq T$) を取得する．総体的対象はカテゴリにおいて，印象を考慮する際に着目すべき対象となる．着目すべき ‘対象’ は，印象についての文書においてより高い関連度を持つといった考えに基づき，オブジェクト名と印象語をクエリとして Web 検索を行い，全て類似オブジェクトに関して，関連度が高い語を総体的対象として取得する．語集合 $Q = \{q_1, q_2, \ldots, q_m\}$ をクエリに AND 検索を行った際に得られる文書集合を D_Q，D_Q に出現する語 w の関連度を $\mathrm{Freq}(D_Q, w)$ と定義すると，総体的対象にあたる語 w は閾値 ϕ を用いて次の条件を満たす語とし，得られた語集合を $T^{aggr}_{o,i}$ とする．

$$\frac{1}{|\{o\} \cup C_o|} \times \sum_{o' \in \{o\} \cup C_o} \mathrm{Freq}(D_{\{o',i\}}, w) > \phi \qquad (1)$$

Web 検索には Bing Search API(注5) を本研究では利用した．語の関連度 Freq には情報検索の分野で広く利用されている $tfidf$ を用いた．これには，tf がよく現れる語であることを表し，idf 値がクエリに対する関連語抽出として役立つためである．tf 値は得られた Web 検索結果に含まれる文書から取得を行い，df 値の算出は，ClueWeb09 Dataset(注6) の日本語データに含まれる 300 万件の文書から生成した値を用いた．

4.2.3 個別的対象の取得

総体的対象に対して，2 つ目の ‘対象’ となる個別的対象の取得方法を説明する．カテゴリ化が出来ないものとして，他のオブジェクトとの比較が行えないような ‘対象’ が挙げられる．例えば原子力発電所が危険といった印象における「放射能物質を拡散する」といった根拠に対し，‘対象’ として放射能物質が得られるが，これは他の類似オブジェクトとの比較が行えない．というのも，発電所といったカテゴリに対する類似オブジェクトには火力発電所や水力発電所などが得られるが，これらの語において放射能物質は全く関係の無いものであり，その ‘対象’ の状態や評価がどうであれ印象を連想する事は難しいためである．直接的に放射能物質という言葉から危険という印象を感じたとしても，それは火力発電所や水力発電所に依るものではない．個別的対象として，他の類似オブジェクトにおいては評価がされない，もしくは評価が出来ないような語の取得を試みる．

個別的対象は入力オブジェクトに特有な語であり，印象を連想させる語である．つまり，入力オブジェクト o における印象に関する文書には出現し，逆に類似オブジェクトにおける印象に関する文書には出現しないような語を取得すれば良い．個別的対象集合を $T^{indi}_{o,i} = \{t^1_{indi}, t^2_{indi}, \ldots\}$ ($T^{indi}_{o,i} \subseteqq T$) と定義すると，$s_u$ は閾値を $\psi(> 0)$ とした時，次の条件を満たすような語 w から $T^{indi}_{o,i}$ を収集する．

$$\mathrm{Freq}(D_{\{o,i\}}, w) - \mathrm{Freq}(D_{\{c,i\}}, w) > \psi \qquad (2)$$

ここで，c は o の類似オブジェクト集合 C_o ($C_o \subseteqq T$) に含まれるオブジェクトである．C_o に含まれる全てのオブジェクトに対し，語 w が式 2 を満たす語であるならば，w は o に特有な語となり，対象として考える事が出来る．

4.3 根拠の収集

前小節で得られた総体的対象集合 $T^{aggr}_{o,i}$ と個別的対象集合 $T^{indi}_{o,i}$ を元に，根拠集合 $E_{o,i}$ を収集する方法を説明する．対象集合には対象 t_u が含まれており，t_u から複数の根拠 $E^u_{o,i}$ が得られる．これは 1 つの ‘対象’ に対して，人々が様々な ‘感じ’ を抱く事に対応する．例として，渋谷が “若者の街” といった印象に対して ‘対象’「ビル」が得られた時,「ビルが多い」のか「ビルが少ない」のか，あるいは「ビルが綺麗」なのか，といった様々な根拠が得られる事に対応する．特に，ここで得られる “多い” “少ない” “綺麗” にあたる語が本研究における ‘対象’ に

(注5)：Bing Search API http://datamarket.azure.com/dataset/bing/search
(注6)：ClueWeb09 Dataset http://lemurproject.org/clueweb09/

対する ‘感じ’s となる．根拠を収集するには，‘対象’ 毎に ‘感じ’ を収集すれば良い．

‘対象’ に対する ‘感じ’ を収集する手法として言語パターンを用いる手法を提案する．用いる言語パターンとして，助詞・助動詞である「が」「な」「は」「を」を用いる．これはある対象の状態を説明する記述として「t が s」「s な t」「t は s」が，対象の動作を説明する記述として「t は s」「t が s」「t を s」といった表現がなされるためである．例として状態を記述する根拠として渋谷が “若者の街” といった印象に対して「ビルが綺麗」といったビルの状態を表す記述を取得するものとする．また，動作を説明する記述とは，原子力発電所が危険に対して「放射線を拡散する」といったような，放射線に対する動作が記述された根拠の事を指す．本研究で取得する対象に対する感じに関して，言語的特徴から，サ変接続名詞・動詞・形容詞のいずれかを取得するものとする．これは口語体で書かれた文脈において評価に値しない誤った語を取得するのを防ぐためである．また，評価に値する語を取得すると同時に，その評価が肯定の意味で用いられているのか，あるいは否定の意味で用いられているのか，といった値についても取得する．これは，否定の意味で用いられている場合には，全く逆の評価がなされていると考えられるため，それらを区別するために用いる．以上から ‘対象’t_u から言語パターンで感じ及び肯定否定の組 (s_u^k, b_u^k) を複数取得し，$E_{o,i}^u = \{(t_u, s_u^1, b_u^1), (t_u, s_u^2, b_u^2), \ldots\}$ $(E_{o,i}^u \subseteqq E)$ とする．

総体的対象集合 $S_{o,i}^{aggr}$ を用いる事で総体的根拠集合 E_{oi}^{aggr} が，個別的対象集合 $S_{o,i}^{indi}$ を用いる事で個別的根拠集合 $E_{o,i}^{indi}$ を収集される．ここから，オブジェクト o の i という印象に対する根拠集合は $E_{o,i} = E_{o,i}^{aggr} \cup E_{o,i}^{indi}$ として扱う．これは，o に関する連想を行う際に，総体的根拠や個別的根拠といった特徴に依らず互いに連想しあう事が出来るためである．原子力発電所が “危険” といった印象に対して，総体的根拠として「事故が多い」が，個別的根拠として「放射線が拡散する」といった根拠が得られる．「事故が多い」といった根拠については，他の類似オブジェクトにおいても評価される根拠であるが，今回原子力発電所の印象についてを考える際，「事故が多い」事から「放射線が拡散する」といった事を連想する事が可能である．よって本研究では本小節で得られた総体的根拠集合及び個別的根拠集合は同等のものとして扱う．

5. 連想に基づく根拠度計算

原子力発電所が危険という印象に対して「放射線が拡散する」「ガンになる」という根拠が得られるが．どちらの方がより根拠らしい根拠と言えるだろうか．本節では，前節で得られた根拠集合 $E_{o,i}$ を入力に，3.3.2 項で定義した根拠度を求める手法について述べる．この際，より多くの人が印象を連想するにあたり参照する根拠が根拠らしい語になるという仮説のもと，根拠度の計算手法を定式化する．

5.1 印象連想モデル

人が印象を抱くには得られた根拠からのみならず，自身が知っている知識に基づき他の根拠を経由した上で印象を抱くことが考えられる．ある根拠を見た上で他の根拠を考慮せずに印象を抱くというモデルは現実的な行動に沿ったものではない．根拠から他の根拠を経由するというのは，根拠から根拠が連想できる事に対応する．つまり，人が印象を抱く過程は図 2 のようになり，より印象を連想出来る根拠を発見するまで他の根拠を連想する過程が踏まれる．人が原子力発電所に “危険” という印象を抱くには，図 3 のような印象連想モデルで表され，一例として，「事故が起こる」という根拠から「放射線が拡散する」「ガンになる」といった根拠を連想した上で印象を抱く事が起こりえる．

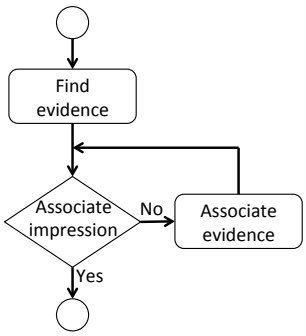

図 2　根拠発見から印象を抱く過程

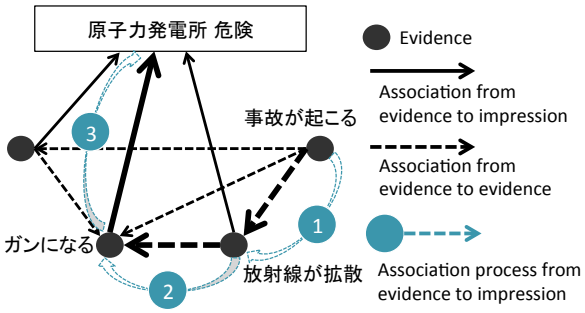

図 3　根拠発見から印象を抱くまでの連想モデル

印象連想モデルから，より多くの人が印象を抱く際に根拠とする情報として，2 つの点に着目できる．1 つ目は特定の根拠を発見した際に，どれ程印象を抱くかどうか，といった指標である．ある根拠から印象を連想しやすいのならば，その根拠を元に印象を抱いたと考える事が出来るためである．また 2 つ目に，他の重要な根拠を連想させる根拠も重要だと考える事が出来る．これは，印象を連想する際に，他の重要な根拠を連想させやすい根拠ならば，そちらの根拠を基にして印象を抱いたと考える事が出来るためである．これら 2 つの点に着目して根拠度を求める式の定式化を図る．

5.2 根拠から印象への連想

根拠度の指標の 1 つ目として，重要な根拠は印象を連想させやすい語である事を用いる．根拠 $e_u (\in E_{o,i})$ がどれ程印象を連想させやすいかの度合いを印象連想度 $\mathrm{Dep}(o, i, e_u)$ と定義する．o に関する文書において e_u がどれ程印象を連想させるかの割合を取れば良いので，$\mathrm{Dep}(o, i, e_u)$ を次式のように定義する．

$$\mathrm{Dep}(o, i, e_u) = \mathrm{Assoc}(\{o, e_u^{st}\}, \{i\}) \tag{3}$$

ここで，集合 A,B に対し，$\mathrm{Assoc}(A, B)$ は連想度を表し，A か

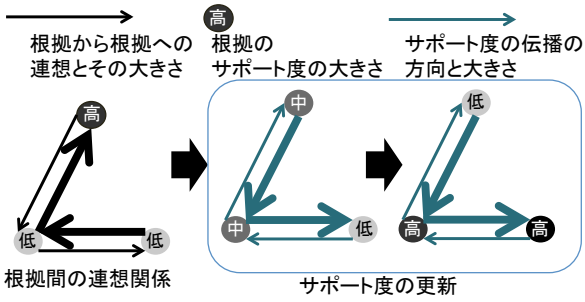

図 4　根拠のサポート度の更新処理

ら B を連想させる度合いを示す．また，e_u^{st} は，根拠 e_u に含まれる対象 t_u，感じ s_u，肯定 or 否定 b_u から，実際に Web 上に現れる文を抽出したものとする．この処理は，4.3 小節で行う根拠の収集の際に同時に行えば良い．e_u が (放射能, 拡散, pos) といった値が入っている際には，e_u^{st} として「放射能が拡散する」といった文が得られる．

連想度 Assoc(A, B) は，A の出現により B が言及されると考えると，A を含むページの内 B を含むページの割合で表す事が出来る．実際に語集合を含む Web ページ数を求めることは非現実的であるため，ここでは語集合をクエリとして AND で繋げた Web 検索を行い，検索結果として得られる検索結果総件数を用いる．$A = \{a_1, a_2, \ldots, a_m\}$，$B = \{b_1, b_2, \ldots, b_n\}$，DF$(a_1 \wedge \cdots \wedge a_m)$ を語 $a_1 \ldots a_m$ をクエリに AND 検索した際の検索結果総件数とした時，連想度を次式で表す．

$$\mathrm{Assoc}(A, B) = \frac{\mathrm{DF}(a_1 \wedge \cdots \wedge a_m \wedge b_1 \wedge \cdots \wedge b_n)}{\mathrm{DF}(a_1 \wedge \cdots \wedge a_m)} \qquad (4)$$

また，DF 値の算出には，4.2.2 項でも用いた ClueWeb09 Dataset の日本語データセットから検索を行った．

5.3　根拠から根拠への連想

根拠度の指標の 2 つ目として，他の重要な根拠を連想しやすい根拠である事を用いる．この度合いを本研究ではサポート度と定義する．これは根拠同士の連想関係を用いて再起的に定義を行う．根拠を 1 つのノード，根拠間の連想関係をエッジと捉えグラフを構築した後に，連想関係の強度によりサポート度を逆向きに流す．また，重要な根拠を連想する事を利用するため，ノードの初期値として 5.2 小節で得られた，依存度を与える．これは topic-sensitive PageRank [4] における枝の向きとは逆に重要度を流す事と等価になる (図 4)．

オブジェクト o の i という印象に関する根拠の連想関係を示すグラフを定義する．ノード集合は $E_{o,i}$ となる．このグラフにおけるエッジは，ノード間の連想関係を示す方向付きエッジとなる．エッジはノード間の全てに存在し，双方向に違う重みを持つ．エッジが持つ重みは根拠から根拠への連想度を表すため，この値を根拠連想度と定義し，オブジェクト o の i という印象に関して e_u から e_v への根拠連想度を EA(o, i, e_u, e_v) と表す．

ノードとなる根拠 e_u が持つサポート度を sup_u と定義し，ノードの初期値として $sup_u^0 = \mathrm{Dep}(o, i, e_u)$ を与える．隣の根拠をどれ程連想させるのか，といった更新処理を繰り返す事で，ノードが持つサポート度を求める事が出来る．sup_u は他の根拠をどれ程連想させるかに基づき，次のような更新式で更新される．

$$sup_u^{j+1} = \alpha \sum_{v \in E_{o,i}} \mathrm{EA}(o, i, e_u, e_v) \times sup_v^j + \frac{1-\alpha}{|E_{o,i}|} \qquad (5)$$

ここで，α は PageRank で用いられているダンピングファクターである．根拠から根拠の連想においても，連想関係に依らず初めに見た根拠とは異なる他の根拠から印象を連想することは，PageRank におけるランダムサーファー同様の考えを用いる事が出来ると考えられる．また，前小節で定義した Assoc を用いて EA$(o, i, e_u, e_v) = \mathrm{Assoc}(\{o, i, e_u^{st}\}, \{e_v^{st}\})$ と定義する．これは，5.3 小節同様，連想前の根拠を含むページのうち，連想先の根拠を含むページの割合を，o と i に関する文書の中から探している．

5.4　根拠度の計算手法

5.1 小節で定義したモデルによると，5.2 小節で定義した印象連想度と 5.3 小節で定義したサポート度の双方が根拠度において重要な役割を持つ．仮に印象連想度のみを持つ根拠であれば，他の根拠を連想する事なく独立して印象を連想させるような根拠となるため，万人に受け入れられるような多様性を持つ根拠とはなりにくい．またサポート度のみを持つ根拠であれば，他の印象を連想させはするものの，実際に印象を連想させるのに充分な根拠となっているとは言えにくい．これは，風が吹けば桶屋が儲かるといったことわざにあるように，結果として様々な根拠を連想させたとしても根拠にはなりえない事を表す．よって双方の値が高い値を持つものが良い根拠として根拠度を定義出来る．

以上の議論から，$dep_u = \mathrm{Dep}(o, i, e_u)$ とした時，根拠度 Evid(o, i, e_u) を印象連想度とサポート度を用いて以下のように定義する．

$$\mathrm{Evid}(o, i, e_u) = \frac{(1+\beta^2) \times dep_u \times sup_u}{dep_u + \beta^2 \times sup_u} \qquad (6)$$

β サポート度に対し印象連想度を何倍重視するかのパラメータである．式 6 は情報検索においてランキングの評価尺度に用いられている F 値 [8] を参考にしたものである．

6.　実　　験

提案手法によって得られる根拠集合及び根拠度についての有効性を評価するため，実験を行った．

6.1　諸　条　件

入力として，オブジェクトと印象を表 1 のように与えた．‘対象’ 集合を取得するための類似オブジェクトには，並列助詞を用いた言語パターンによる取得 [13] を行った後に，得られた類似オブジェクトの中から人手で選定を行った．

実験における条件やパラメータとして，対象収集のための Web 検索には検索結果上位 100 件を利用，また，個別的対象集合や総体的対象集合の数が各々100 件になるよう式 1 の ϕ と式 2 の ψ を設定した．また，サポート度を求める際の更新に用いるダンピングファクターには，PageRank で有用とされる 0.85 及び ランダムサーファを許容しない 1 の 2 つの値を用いた．

4 節で定義した根拠の収集を得られた対象を元に Web 検索

表 1　評価実験における入力値及び類似オブジェクト集合

オブジェクト	印象値	類似オブジェクト
猫	マイペース	犬
犬	従順	猫
クマ	凶暴	トラ，イノシシ
原子力発電所	危険	火力発電所，水力発電所
渋谷	若者の街	新宿，品川
秋葉原	電気街	神田，日本橋
イチロー	アベレージヒッター	松井秀喜，野茂英雄
王貞治	ホームランバッター	長嶋茂雄，松井秀喜
オタク	キモい	マニア
東大生	賢い	京大生，阪大生
冬	寒い	春，夏，秋
メール	便利	電話，手紙
野菜	健康に良い	果物，肉

結果上位 100 件の文書から探索を行う事で，例として猫が “マイペース” からは計 234 件の根拠を，原子力発電所が “危険” という印象からは計 470 件の根拠を発見した．

6.2 評価実験

本提案手法の有効性を示すため，得られた根拠集合を基に複数パターンの根拠度計算手法を適用して根拠度の高い根拠上位 n 件に対する比較を行う．まず初めに，オブジェクト o と印象 i を AND 検索にて得られた Web 検索結果に含まれる文書中に出現する頻度順でランキングを行った根拠をベースラインとして取得した．これはより重要と考えられる根拠は関係する文書中に何度も現れるであろうことに基づいている．提案手法における根拠度には 5.2 小節で定義した根拠の印象連想度と 5.3 小節で定義したサポート度が存在するため，これらのいずれかのみを利用した手法を別に 2 つ用いる事とする．また，式 6 で定義した根拠度にはサポート度に対し印象連想度を何倍重視するかのパラメータ β が含まれる．よって，これらの値を変えたものを 3 つ用意し，計 6 個の手法で根拠度の算出を行う．用いた手法は以下のように表す．

- freq：根拠の出現頻度に基づくベースライン手法
- dep：印象連想度のみを根拠度として用いる手法
- sup：サポート度のみを根拠度として用いる手法
- even：$\beta = 1$ とし，重みを等しく取る手法
- dep+：サポート度に対し印象連想度に 2 倍の重みを置く手法
- sup+：印象連想度に対しサポート度に 2 倍の重みを置く手法

評価の際，初めに得られた根拠集合に各手法を用いて，根拠度の高い上位 20 件の根拠をプーリングする．その後に，被験者に重複を除いた根拠集合をランダムに表示した上で，「印象を連想させる語であるか」といった問いに対して各根拠毎に Yes/No で回答を受け付ける．得られた根拠の評価結果を基に正解セットを作成し，各手法の有効性を評価する．

今回行った評価実験は被験者を 20 代の男性 3 名とし，各手法を適応した上でプーリングした結果の根拠集合，例として猫が “マイペース” に対して 57 件，原子力発電所が “危険” に対して 63 件を評価対象として用いた．本研究における根拠らしい根拠とは，より多くの人が同様に考える根拠の事であるため，3 名の被験者中 2 名が Yes と答えたものを正解セットとして用いた．例として原子力発電所が “危険” といった印象における正解セットは表 2 のようになった．なお，ここでの根拠文とは根拠に対応する Web 上に実際に出現していた文のことである．

表 2　原子力発電所が危険という印象における正解セット

根拠文	
ウラニウムが流れ出す	浜岡原発が危険
建屋が爆発	事故は起きる
深刻なセシウム	事故がある
リスクをはらむ	事故がおきる
事故が起こる	セシウムが露出
放射能を帯びる	放射線を放つ
危険な放射線	放射能をふくむ
セシウムが検出	ウラニウムが飛散
浜岡原発が爆発	危険な原発
放射能をあびる	

表 4　全てのタスクにおける適合率 (P@k) の平均

	d = 0.85		d = 1	
Method	P@10	P@20	P@10	P@20
freq	0.2333	0.2083	0.2333	0.2083
dep	0.4500	0.4000	0.4500	0.4000
sup	0.1667	0.1333	0.1500	0.1250
even	0.4667	0.4167	0.4500	0.4000
dep+	**0.5000**	**0.4333**	**0.5000**	**0.4417**
sup+	0.3833	0.3167	0.3667	0.3083

得られた正解セットを基に，各手法の有効性を適合率を用いて評価するとタスク毎の結果は表 3 のようになった．タスクに依り提案手法が良い結果を出しているものが多い一方，正解セットの数が少ないことから差が見られない物，ベースラインより悪い結果を出しているものもミられた．差が見られないものに関して，今回収集した根拠集合自体が誤っている事，収集の際のパラメータが小さい事により根拠として扱うべき ‘対象’ が得られなかった事が考えられる．悪い結果を出したタスクにおいては，印象連想度が高い根拠が実際の人が思う印象連想度とは異なる値になっていたと考えられる．提案手法においては文書における共起のみに着目しており，今後検討の余地があると考えられる．

手法間の比較を行うため，全てのタスクにおける適合率の平均を取ると表 3 の結果が得られた．ここから，ベースラインに対して提案手法が良い根拠を発見出来た事が分かる．また，印象連想度とサポート度の片方のみに着目するよりか，双方の値が高い値を持つ手法が良い結果となった．特に，印象連想度に大きな重みを置いた手法が最も良い結果となった．ここから，より印象を連想させる根拠に着目するのみならず，他の根拠を連想させるような特徴も根拠を発見する上で役立っていたのではと考えらえる．このような結果は，人が他の根拠を連想した上で印象を連想するよりかは，直接的に根拠から印象を抱く事が多いといったモデルに当てはめる事が出来る．このような行動にも何ら不自然な点はなく，その点からも今回提案した手法の有効性が言える．

表 3　タスク毎の印象を連想させる根拠上位 k 件の適合率 P@k ($d = 0.85$)

	猫 マイペース		原子力発電所 危険		クマ 凶暴		冬 寒い		犬 従順		渋谷 若者の街		イチロー アベレージヒッター	
Method	P@10	P@20	P@10	P@20	P@10	P@20	P@10	P@20	P@10	P@20	P@10	P@20	P@10	P@20
freq	0.00	0.05	0.20	0.25	0.00	0.15	0.20	0.15	0.10	0.05	0.20	0.15	**0.20**	**0.10**
dep	**0.30**	**0.30**	0.40	0.35	**0.20**	**0.35**	0.30	0.25	**0.30**	**0.20**	**0.80**	**0.55**	0.10	0.05
sup	0.00	0.00	0.50	0.40	0.00	0.05	0.20	0.10	0.00	0.00	0.00	0.00	0.10	0.05
even	**0.30**	**0.30**	**0.70**	**0.55**	**0.20**	**0.35**	**0.50**	**0.35**	**0.30**	**0.20**	0.30	0.35	0.10	0.05
dep+	**0.30**	**0.30**	0.50	0.50	**0.20**	**0.35**	**0.50**	**0.35**	**0.30**	**0.20**	0.70	0.50	0.10	0.05
sup+	**0.30**	0.25	0.50	0.40	**0.20**	0.25	0.30	0.25	**0.30**	**0.20**	0.20	0.10	0.10	0.05

評価実験から 2 つの課題が考えられる．1 つ目は根拠を収集する際のパラメータをどのように調節するかといった課題である．これらの値が大きすぎると得られる根拠が少なくなってしまい，本来根拠となるべき根拠が収集できない．しかし小さくしすぎた場合にもノイズとなる根拠が多く収集されてしまうため，この際，入力として与えられたオブジェクトと印象に応じてどの程度根拠を収集すべきかを変化させるといった事が考えられる．また，2 つ目に印象連想度とサポート度の取る重みをどのようにするのか，といった課題がある．今回行った評価実験における総合平均からは印象連想度に大きな重みを付けた手法が良い結果となったが，重みを変更することでさらに良い根拠を発見出来るのではないかと考えられる．

7. まとめと今後の課題

本稿では，入力オブジェクトのある印象に対する根拠の発見と，発見した根拠の根拠らしさを評価する手法を提案した．根拠の発見には社会心理学で提言された連続体モデルに着目し，カテゴリ化と個別化の双方において評価されるであろう ‘対象’ から収集を行った．また，‘対象’ に対しどのような評価がなされているかについて，言語パターンを用いて ‘感じ’ を取得する手法を提案した．収集した根拠の根拠らしさの評価には，人の印象を抱く過程をモデル化し，モデルに基づき根拠が印象をどれ程連想するかという印象連想度，根拠が他の重要な根拠をどれ程連想するかというサポート度の双方を用いた根拠度の評価手法を提案した．また，提案した根拠度の評価手法が有効であるかを確かめるための評価実験を行い，頻度に基づく手法に比べ，より多くの人が根拠と考える根拠を取得出来た事を示した．

また，根拠の出現確率を等価なものとして扱ったため，これを変化させるような手法も考えられる．本来，人がある根拠を発見する確率というのは等価にはならない．他の人がより多く言及しているような根拠であればその根拠は多くの人に発見されやすくなり，結果としてその根拠を用いて印象を連想する人が増える．また，別の課題として今回得られた根拠などの信憑性を考えるといった事が考えられる．山本らは Web ページの信憑性に関する研究 [11] も行っており，このような手法を適用する事で，発見した根拠の信憑性を計る事が出来る．ここから誤った根拠から印象が連想されている，といった事象の発見にも本研究は貢献出来るのではないかと考えられる．

謝辞　本研究の一部は，文科省科研費基盤 (A)「ウエブ検索の意図検出と多元的検索意図指標にもとづく検索方式の研究」(24240013，研究代表者：田中克己)，文部科学省科学研究費補助金（課題番号 24240013，24680008）によるものです．ここに記して謝意を表します．

文　献

[1] N H Anderson. Averaging versus adding as a stimulus-combination rule in impression formation. *J Exp Psychol*, Vol. 70, No. 4, pp. 394–400, 1965.

[2] Solomon Asch. Forming impressions of personality. *Journal of Abnormal and Social Psychology*, Vol. 41, pp. 258–290, 1949.

[3] S. T. Fiske and S. L. Neuberg. A continuum model of impression formation, from category-based to individuating processes: Influence of formation and motivation on attention and interpretation. *Advances in experimental social psychology. New York: Academic Press.*, Vol. 23, pp. 1–74, 1990.

[4] Taher H. Haveliwala. Topic-sensitive pagerank. In *Proceedings of the 11th International Conference on World Wide Web*, WWW '02, pp. 517–526, New York, NY, USA, 2002. ACM.

[5] Jon M. Kleinberg. Authoritative sources in a hyperlinked environment. *JOURNAL OF THE ACM*, Vol. 46, No. 5, pp. 604–632, 1999.

[6] R. Mihalcea and P. Tarau. TextRank: Bringing order into texts. In *Proceedings of EMNLP-04and the 2004 Conference on Empirical Methods in Natural Language Processing*, July 2004.

[7] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The pagerank citation ranking: Bringing order to the web. Technical Report 1999-66, Stanford InfoLab, November 1999. Previous number = SIDL-WP-1999-0120.

[8] C. J. Van Rijsbergen. *Information Retrieval*. Butterworth-Heinemann, Newton, MA, USA, 2nd edition, 1979.

[9] T.K. Srull and R.S. Wyer. *A Dual process model of impression formation*. Advances in social cognition. L. Erlbaum Associates, 1988.

[10] Peter D. Turney. Expressing implicit semantic relations without supervision. In *In Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL-2006*, pp. 313–320, 2006.

[11] Yusuke Yamamoto. Disputed sentence suggestion towards credibility-oriented web search. In Quan Z. Sheng, Guoren Wang, Christian S. Jensen, and Guandong Xu, editors, *APWeb*, Vol. 7235 of *Lecture Notes in Computer Science*, pp. 34–45. Springer, 2012.

[12] 稲川雅之, 大島裕明, 小山聡, 田中克己. Web からの語集合間の特定関係の抽出とその可視化. 日本データベース学会論文誌, Vol. 7, No. 1, pp. 175–180, 2008.

[13] 大島裕明, 小山聡, 田中克己. Web 検索エンジンのインデックスを用いた同位語とそのコンテキストの発見. 情報処理学会論文誌. データベース, Vol. 47, No. 19, pp. 98–112, 2006.