F-12

Web からの知識抽出による
閲覧ページの動的なマルチファセット生成

川野 悠†　　大島 裕明†　　田中 克己†
Yu Kawano　　Hiroaki Ohshima　　Katsumi Tanaka

1．はじめに

近年のインターネットの発達により，我々は Web ページから情報を得る機会が多くなった．もし興味のある情報に関する Web ページを見つけたいのなら，検索エンジンを利用することで膨大な数の Web ページから検索することも可能である．興味のある情報に関する Web ページを閲覧している場合，ユーザはその情報に関連のある別の Web ページを閲覧したいという欲求に駆られることがしばしばある．例えば，図 1 のように C 社のプリンタに関する Web ページを閲覧しているユーザを考える．もしそのユーザが C 社の製品に興味を持っていてその Web ページを閲覧しているのであれば，C 社のカメラなど他製品に関する情報を知りたいのではないかと考えられる．一方で，ユーザがプリンタに興味を持っているという状態でその Web ページを閲覧しているのならば，エプソンや hp などの他のメーカーのプリンタに関する情報を求めていると考えられる．このように，1 つの Web ページに複数のトピックを含むようなページは，ユーザによって違った観点から閲覧されているといえる．

このような Web ページに対して関連ページをユーザに提示する場合，単に関連ページをリスト形式で提示するだけでは不十分であり，Web ページを閲覧するための観点毎に関連ページを提示する必要があると考えられる．観点とは上の例を用いると「メーカー」や「製品」にあたるものであり，「C 社」や「プリンタ」という語をより一般化した語ということができる．

そこで本稿では，閲覧ページに含まれるトピック毎に動的に複数のファセットを生成する手法を提案する．ファセットとは物事を見る側面のことで，本稿では閲覧ページの持つ観点のことを指す．この手法を用いることで，閲覧ページのファセット毎に関連ページを分類して提示することが可能となる．

2．閲覧ページのファセット生成

本節では生成するファセットの定義を与えた後，閲覧ページから複数のファセットを生成する手順について述べる．

2.1　ファセットの定義

電化製品を分類する場合，「メーカー」という観点から「パナソニック」や「ソニー」などの各メーカーに分類することができる．しかし，「製品名」という観点から「テレビ」や「エアコン」などの製品別に分類することも可能である．このように，分類には観点と分類されたオブジェクトに付けるラベルが必要であり，本稿では分類の観点をファセット名，分類のラベルに当たるものをファセット値と定義する．

図1　閲覧ページ例
（http://cweb.canon.jp/pixus/index.htm より引用）

本稿で生成するファセットは，一つのファセット名と複数のファセット値からなる組である．先ほどの例であれば，ファセット名が「メーカー」，ファセット値がそれぞれのメーカー名となるファセットと，ファセット名が「製品」，ファセット値がそれぞれの製品名となるファセットが生成される．

2.2　マルチファセット生成の手順

本稿では入力として閲覧ページが与えられ，そのページがどのようなファセットのどのファセット値に分類されるのかを推定する．手順としては，まず閲覧ページに含まれる複数のトピックを推定する．各トピックは分類先のファセット値に相当する．次に，トピックごとにそのトピックの同位語を発見する．発見された同位語はトピックとは異なる別のファセット値に相当する．最後に，発見された複数のファセット値の軸となるファセット名を発見する．この操作をトピックごとに行い，複数のファセットを生成することができる．

例えば，図 1 のような C 社のプリンタに関する Web ページからは「C 社」と「プリンタ」という 2 つのトピックが推定され，ファセット名としてそれぞれ「メーカー」と「製品」，ファセット値として「エプソン,hp」と「デジカメ,複合機」などが得られると考えられる．

3. 検索エンジンを用いたマルチファセット生成手法

本節では検索エンジンを用いて閲覧ページのマルチファセットを動的に生成する手法を提案する．本手法は閲覧ページに含まれるトピックの推定，トピックの同位語

†京都大学大学院　情報学研究科　社会情報学専攻

発見，同位語集合からの上位語発見の 3 つのフェーズから構成される．

3.1 閲覧ページに含まれるトピックの推定

閲覧ページのファセットを生成するために，まずそのページに含まれるトピックを推定し抽出する．ここで，トピックとはそのページの主題を表す語として定義する．例えば閲覧ページが「C 社のプリンタに関するページ」であれば，「C 社」や「プリンタ」といった語はそのページを表現するのに欠かせない語であり，そのページの主題であるといえる．抽出したトピックはそれぞれ，生成するファセット中のファセット値に相当する．

トピックはそのページの主題であるので，本文中に頻出したり，タイトルや見出しに出現すると考えられる．また Web ページには一般的に複数のトピックが含まれている．そこで，ある語 t の主題度 S_t を以下の式で定義し，主題度が高い上位 n 件をトピックとして抽出する．

$$S_t = \sum_{tag \in Tag} \alpha_{tag} \cdot tf_{tag}$$

ここで，html 内に含まれるタグ集合を Tag とし，その中に含まれる任意のタグ tag に対する重みを α_{tag}，t の出現頻度を tf_{tag} とする．一般的にはタイトルに出現する語と本文中に出現する語が同頻度であれば，タイトルに出現する語のほうが主題を表していることが多い．そこで，例えば<title>タグの重みを 10，<body>タグの重みを 1 とすることで，タグの重みを考慮した主題度を算出できる．

3.2 トピックの同位語発見

抽出したそれぞれのトピックに対してファセットを生成するために，まずトピックの同位語を発見する．これはトピックとは異なる，別のファセット値を発見するためのアプローチであり，分類においてファセット値同士は同位語であることが望ましいという仮定に基づいている．

同位語の発見には大島ら[1]の手法を改良した手法を用いる．大島らの手法は，ある語とその同位語は互いを助詞の「や」で接続したフレーズとして用いられやすいことに着目している．例えば，「C 社」とその同位語と考えられる「エプソン」は「C 社やエプソン」または「エプソンや C 社」といったフレーズで使われることが多いと考えられる．そこで，まず「C 社」の前後それぞれに「や」を接続し，「や C 社」と「C 社や」という 2 つの構文パターンを作成する．前者の構文パターンを前パターン，後者を後パターンと呼ぶことにする．次に，2 つの構文パターンそれぞれをクエリとして Web 検索を行い，タイトルとスニペットを得る．もし前パターンの検索によるタイトルやスニペット中に「エプソンや C 社のような」という記述があり，かつ後パターンの検索によるタイトルやスニペット中に「プリンタは C 社やエプソンが」という記述があれば，「エプソン」を同位語として抽出する．このように，前パターンの直後に出現する語と後パターンの直前に出現する語が一致していれば，その語を同位語として抽出する．

しかし，この手法は両方の構文パターンを含む表現に現れた語のみを同位語とみなすため，求める語によっては再現率が低い場合が存在する．例えば，「C 社」という語の前パターンに一致する表現として「hp や C 社」が得られても，後パターンに対応する表現として「C 社や hp」が得られるとは限らず，もしその表現が Web 検索から得られなければ「hp」は同位語として抽出されることはない．次のフェーズで行う上位語発見は，同位語の数が多い方が有利に働くので，本稿ではこの手法の「両方の構文パターンに一致する」という部分を緩和した手法を提案する．

この手法は同位語に対する同位語は，元の語の同位語である可能性があるという考えに基づいている．例えば「巨人」というプロ野球球団の同位語として「ソフトバンク」が得られたとする．このとき「阪神」という同位語は得られなかったが，「ソフトバンク」の同位語を再度発見することで「阪神」という語が得られる可能性がある．この場合，「巨人」の同位語として「阪神」も認めることが考えられる．しかし，「ソフトバンク」は携帯電話事業者と見ることもでき，同位語として「ドコモ」という語が得られると，「巨人」の同位語として「ドコモ」も認めることになってしまう．

そこで，「巨人」の同位語を発見する際に用いた，Web 検索によって得られた表現を利用する．「巨人」と「ドコモ」は同位関係にないので，「巨人」の同位語を発見する際に用いた表現の中で，「ドコモ」という語が出現する可能性は極めて低い．しかし，「巨人」と「阪神」は実際には同位関係にあるので，「巨人」の同位語発見に用いられる表現のうち，両方の構文パターンには含まれないが，前または後のいずれかのパターンを含む表現には「阪神」という語が含まれる可能性がある．例えば前パターンにおいて，「阪神や巨人にまつわる」という表現が得られても，後パターンには「巨人や阪神」というフレーズを含む表現が見つからなかった場合，元の手法では「阪神」を同位語として抽出できないが，我々の緩和手法では「阪神」を同位語として抽出することができる．このように「両方の構文パターン」という条件を一部緩和することで，より多くの同位語を発見することができると考えられる．

ここからは条件を緩和した同位語発見手法を用いて，トピック t の同位語を発見する手順を述べる．定義として，ある語 t に対して前パターンを含む表現のみから得られる語集合を $Pre(t)$，後パターンを含む表現のみから得られる語集合を $Post(t)$，前パターンを含む表現と後パターンを含む表現から共通して得られる語集合を $All(t)$ と表すことにする．

(1)まずあるトピック t に対し語集合 $All(t)$ を得る．
　$w \in All(t)$ は t の同位語とみなす．
(2)同位語 $w \in All(t)$ に対し $All(w)$ を得る．
(3) $All(w)$ に含まれる語 w' が、$w' \in Pre(t)$ または $w' \in Post(t)$ であれば，w' も t の同位語とみなす．
(4)全ての w に対して、(3)を行う．

この結果得られた同位語集合を C_t，トピック t も含めたファセット値の集合を $FValue$ と表す．それぞれのトピックに対して同位語集合の発見を行う．

3.3 トピックの上位語候補の発見

それぞれのトピックに対し，上位語候補を発見する．得られた上位語はファセット名に相当し，ファセットが生成される．

本稿では複数の同位語集合から上位語を発見する手法を提案する．基本的なアイディアは同位語発見の際に用いた考え方と同じである．大島らの同位語発見手法を一般化し，ある語と特定の関係にある語（以下関連語と呼ぶ）を発見する手法[2]を改良することで実現する．

同位語発見手法では，元の語の前後それぞれに「や」を接続することで構文パターンを作っていたが，一般化した手法では前後に接続する語を変えることで，それに応じた関連語を発見することができる．例えば，上位語を発見したい場合，元の語の前に接続するパターンとして「などの」，後に接続するパターンとして「といえば」を用いることで，上位語の発見に適用できると考えられる．しかし同位語発見手法と同様の問題点として，両方の構文パターンを含む表現に現れた語のみを関連語抽出の対象とするため再現率が低いことが挙げられる．また同位語の場合は「や」を接続することで高い抽出精度を誇っているが，他の関連語では必ずしも抽出精度の高い構文パターンを作ることはできるとは限らない．特に上位語に関して言えば様々なパターンを用いて抽出を試みたが，精度が「や」に匹敵するパターンを見つけることは出来なかった．

そこで関連語発見手法の抽出条件を緩和することで，より多くの関連語を抽出する手法を提案する．考え方は前節の同位語発見の拡張手法と同じであり，同位語の関連語は元の語の関連語である可能性が高いことに着目する．例えば入力として「阪神」とその同位語である「巨人」が与えられ，「巨人」の上位語として「プロ野球球団」という語が得られたとする．この場合「阪神」の上位語発見に用いられる表現のうち，両方の構文パターンには含まれないが，前または後のいずれかのパターンを含む表現に「プロ野球球団」という語が含まれるならば，「阪神」の上位語として「プロ野球球団」という語も認めることにする．ここで前後いずれかの構文パターンを含むという条件を課しているのは，抽出の精度を維持するためである．この関連語発見手法は再現率が低いことに加えて，同位語発見手法に比べて精度が劣る場合があることは先ほど指摘した．このような条件を課すことで，同位関係にある 2 語の共通の上位語を抽出することが可能となり，精度の向上も期待される．この手法を採用することで精度を保ちつつ，より多くの関連語を発見することができると考えられる．

ここからは条件を緩和した関連語発見手法を用いて，トピック t の上位語を発見する手順を述べる．

(1) まずトピック t のファセット値集合 $FValue_t$ に含まれる任意のファセット値 fv に対し，改良した関連語発見手法を用いて，前に接続するパターンだけに出現した語集合 $Pre(fv)$，後に接続するパターンだけに出現した語集合 $Post(fv)$，両方のパターンに出現した語集合 $All(fv)$ を抽出する．

(2) ファセット値 fv がトピック t の場合，$h \in All(t)$ は t の上位語とみなす．

(3) ファセット値 fv がトピック t の同位語 c の場合，$All(c)$ に含まれる語 h' が，$h' \in Pre(t)$ または $h' \in Post(t)$ であれば，h' も t の同位語とみなす．

(4) 全ての同位語 c に対して(3)を行う．

この結果得られた上位語の集合はファセット名の候補となり，$FName(t)$ と表す．

3.4 上位語候補のランキング

ファセットを生成するには，ファセット名の候補となる上位語の集合 $FName(t)$ から最適な上位語を 1 つ選択する必要がある．そこで，ファセット値として発見した同位語とファセット名として発見した上位語にそれぞれ評価値を設けることで，最も高い評価値となった上位語をファセット名として最適な上位語と判断し，$FName(t)$ から 1 つ選択する．

まずトピック t に対する同位語 c の評価値 $\mathrm{CValue}(t,c)$ を以下の式で求める．

$$\mathrm{CValue}(t,c) = \mathrm{Bi}(t,c) + \sum_{w \in All(t)} \frac{\beta \cdot \mathrm{Bi}(w,c)}{\mathrm{Rank}(t,w)}$$

この評価式の $\mathrm{Bi}(t,c)$ はどちらのパターンからも共通して得られる同位語集合 $All(c)$ のみに与えられる値であり，残りの項は拡張した同位語発見手法によって新たに得られる同位語に対して与えられる値に寄与する．

$\mathrm{Bi}(t,c)$ は大島らの手法によって得られた同位語に対して与えられる値であり，前と後に接続するパターンそれぞれに出現した c の数を相乗平均したものを採用する．例えば，「C 社」の同位語である「エプソン」が前パターンに 2 回、後パターンに 8 回出現すれば，そのときの $\mathrm{Bi}(t,c)$ は 4 となる．

残りの項は拡張した同位語発見手法によって新たに得られる同位語のための項である．$\mathrm{Rank}(t,w)$ は同位語 $w \in All(t)$ を $\mathrm{Bi}(t,w)$ の大きい順に並べたときの順位であり，t に対する w の同位語らしさを表した順位だと考えることができる．ここで，新たな提案手法によって得られた同位語 c の評価値を決めるのだが，w の同位語らしさが大きいほど c の評価値を大きくするために，$\mathrm{Bi}(w,c)$ と $\mathrm{Rank}(t,w)$ の除算で表している．また β は $0 < \beta < 1$ を満たす実数であり，$\mathrm{Bi}(t,c)$ とのバランスを考慮するものである．

最終的に得られた同位語 $c \in C_t$ を求めた評価値の大きい順に並べ変える．並び変えられた同位語 c の順位を $\mathrm{CRank}(c)$ で表す．

表 1　就職活動における面接のマナーに関するページのファセット

トピック	ファセット値@5	ファセット名@3	ファセット名@1
面接	職務経歴書，履歴書，エントリーシート，電話，作文	就職活動，緊張，就職	就職活動
マナー	ルール，作法，礼儀，しきたり，エチケット	基本，常識，研修	基本

表 2　キャノンの PIXUS のページのファセット

トピック	ファセット値@5	ファセット名@3	ファセット名@1
ピクサス	キャノン，カラリオ，エプソン，イクシ，複合機	プリンター，プリンタ，マー	プリンター
キャノン	ニコン，トヨタ，ソニー，ホンダ，富士フィルム	カメラ，カメ，デジカメ	カメラ

次に，トピック t に対する上位語 h の評価値 $\mathrm{HValue}(t,h)$ を以下の式で求める．

$$\mathrm{HValue}(t,h)=\mathrm{Bi}(t,h)+\sum_{c\in C_t}\frac{\gamma\cdot\mathrm{Bi}(c,h)}{\mathrm{CRank}(t,c)}$$

基本的なアイディアは同位語の評価値と同じで，t に対して関連語発見手法を用いて得られた $All(t)$ から与えられる成分 $\mathrm{Bi}(t,h)$ と t の同位語 c を経由して発見された上位語 h に与えられる成分 $\frac{\gamma\cdot\mathrm{Bi}(c,h)}{\mathrm{CRank}(t,c)}$ の和から成り立っている．

$\mathrm{Bi}(t,h)$ の計算は同位語の場合と全く同じなので省略する．残りの項は同位語 c の h に対する上位語らしさを c の t に対する同位語としての順位で割ったものであり，c の順位が下位なほどこの成分の値は小さくなる．γ は $0<\gamma<1$ を満たす実数であり，この成分の影響を調節するためのパラメータである．

上位語の集合として得られたファセット名候補集合 $FNam(t)$ の中から $\mathrm{HValue}(t,h)$ の最も大きい上位語をファセット名とする．

以上から，1 つのファセット名と複数のファセット値からなるファセットを生成することができ，この操作を全てのトピックについて行うことで，動的にマルチファセットを生成することができる．

4．実験

本稿で提案した Web ページからの動的なマルチファセット生成手法を実装し，実際の Web ページにおいてどのようなトピックおよびファセットが生成されるのか実験した．

4.1　実験設定

今回の実験では，タグに含まれるテキストから形態素解析器 Mecab を用いて名詞または名詞句と推定される語を抽出し，その中から閲覧ページに含まれるトピックの推定を行った．抽出するトピックは主題語の高い上位 2 つとした．同位語に関しては評価値の高い上位 5 件をファセット値として抽出して上位語発見に使用した．同位語，上位語を発見する際の Web 検索には Yahoo! Japan によって提供されている API を使い，検索結果上位 50 件のタイトルとスニペットを利用した．

4.2　実験結果と考察

今回は就職活動における面接のマナーに関するページ[3]とキャノンの PIXUS のページ[4]を対象として実験を行い，それぞれ表 1，表 2 のような結果となった．

各表はそのページで抽出されたトピック，ファセット値の上位 5 件，ファセット名の上位 3 件を示している．表 1 に関しては，このページにふさわしいトピックが抽出されている．また「面接」というトピックのファセットを見ると，ファセット名「就職活動」に対して「履歴書」や「エントリーシート」というファセット値が抽出されており，望ましい結果であるといえる．トピック「マナー」ではファセット名として「基本」が抽出されており，「基本」や「情報」といった単語は Web 上の記述としてよく現れるが，ファセット名として相応しくないので，ストップワードなどで予め取り除いておく必要がある．表 2 に関しては，トピックはうまく抽出できていると思われる．しかし，トピック「ピクサス」のファセット値を見ると，「キャノン」や「エプソン」といったプリンタメーカーが混ざっている．これは「ピクサス」と「キャノン」に強い依存関係があり，「ピクサス」という語の周りには「キャノン」が出現しやすいため，誤って抽出されたものと考えられる．

これらの結果から，表 1 の場合のように本手法が有効なページも存在するが，表 2 のように語同士に強い相関関係がある場合は同位語，上位語の抽出精度が低下する恐れがあるため，何らかの工夫が必要である．

5. まとめと今後の課題

本稿では，閲覧ページに含まれるトピック毎に動的に複数のファセットを生成する手法を提案した．トピックの同位語と上位語を発見することでファセットを生成し，上位語発見の再現率を向上させるために既存の同位語発見手法と上位語発見手法の拡張手法を提案した．今回は提案手法を用いて数件のページのマルチファセットを生成したが，もっと多くの Web ページに対し定量的な指標を用いて評価する必要がある．

今後は提案手法を利用して，閲覧ページのファセット毎に関連ページを分類して提示するシステムを実現させるために研究に取り組んでいく所存である．

謝辞　本研究の一部は，京都大学 GCOE プログラム「知識循環社会のための情報学教育研究拠点」，および，文部科学省科学研究費補助金特定領域研究「情報爆発時代に向けた新しい IT 基盤技術の研究」，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），および，文部科学省科学研究費補助金若手研究（B）「オンデマンド利用を目的とする Web からの知識発見に関する研究」（研究代表者：大島裕明，課題番号：21700105），および，独立行政法人情報通信研究機構の高度通信・放送研究開発委託研究「電気通信サービスにおける情報信憑性検証技術に関する研究開発 課題ア Web コンテンツ分析技術」（研究代表者：田中克己）によるものです．ここに記して謝意を表します．

文　献

[1] 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(トランザクション) データベース, **47**, pp. 98-112 (2006)

[2] 大島裕明, 田中克己：“両方向構文パターンを用いた Web 検索エンジンからの高速関連語発見手法", 情報処理学会研究報告, **88**,pp. 37-42 (2008)

[3]http://www.easy-mensetsu.com/8_point/03basic_manner.html

[4] http://cweb.canon.jp/pixus/index.html