Vol.2010-DBS-151 No.33
2010/11/13

Web上での概念間の共通性に基づく比較可能性の判定

岡田 崇[†1] 湯本高行[†1]
新居 学[†1] 高橋 豊[†1]

本論文では，Web 上での概念間の共通性に基づいた比較可能性の判定を行う手法を提案する．比較可能性とは概念が持つ重要な特徴が共通しているかどうかを表すものであり，似ているかどうかを表す指標である類似性とは異なる．概念は一般的な言葉を複数用いて説明できる語であり，その説明には他の概念や説明語を用いて表現される．説明語とは一般的に用いられる言葉のことで，それ以上の説明が不要な語である．本研究で扱う概念は商品や機能などに限定している．比較可能性の判定には，概念と説明語から生成した重み付き有向グラフを用いて行う．具体的には，概念の持つ特徴などを説明している文から重要な語を抽出し，それらの集合の関係から判定を行う．提案手法の有用性を評価するために，評価実験を行った．実験により，比較可能な概念間において，共通する重要な特徴が抽出できていることがわかった．

Judging Comparability between Concepts based on commonality on the Web

Takashi OKADA,[†1] Takayuki YUMOTO,[†1]
Manabu NII[†1] and Yutaka TAKAHASHI[†1]

In this paper, we propose a method to judge comparability between concepts by commonality on the Web. We regard two concepts are comparable when they have commonality. So, comparability is different from similarity. The concept is explained with other concepts and explanation words. Our main targets of the concepts are products and functions. The explanation word is a general term. We extract keywords from sentences explaining features of concepts. Then, we judge the comparability using similarity of them. We evaluated our method. From the result, we found that our method can extract important common feature between the concepts.

1. はじめに

近年，インターネット上で商品やサービスを購入することが増えている．時間や場所を問わずに商品を購入することができるネットショッピングの利用者数は年々増加している．商品についての情報を調べるための手段の一つとして検索エンジンを利用することが挙げられる．検索エンジンを利用することで，インターネット上に公開されている情報をキーワードを入力することによって検索することができる．ユーザのニーズに合った商品が見つかればその商品を購入することになるが，ユーザは可能な限り多くの商品の中から最適な商品を選択したいと考える．しかし，入力したキーワードと比較すべき商品を検索することはユーザ自身がある程度知識を持っていないと難しい．なぜなら，商品名や比較すべきポイントが分からなければ具体的なキーワードを入力することができないからである．

そこで本研究では，ユーザが入力した概念と比較可能な概念，および各々が持つ重要な特徴や性能を提示することで概念間の比較が行いやすくなり，ユーザの負担を軽減することができると考えている．概念間に共通する重要な特徴が存在すれば比較ができると考え，これを比較可能性と定義する．本稿では，概念間の比較可能性を判定する手法を提案する．比較可能性とは概念間に共通する重要な特徴が存在するかどうかを表すものであり，似ているかどうかを表す指標である類似性とは異なる．ここで，概念は一般的な言葉を複数用いて説明できる語であり，他の概念や説明語を用いて表現されるものとする．説明語とは一般的に用いられる言葉のことで，それ以上の説明が不要な語である．本研究で扱う概念は商品や機能などに限定している．概念の持つ特徴などを説明している文から重要な語を抽出し，それらの集合の関係から判定を行う．具体的なイメージ図を図 1 に示す．「プラズマクラスター」と「ナノイー」はイオンを利用した空気清浄機能であり，浮遊菌やアレル物質，臭いの除去などを行うという点で比較可能であるといえる．

以下，2 節で関連研究，3 節で概念間の比較可能性の判定手法について述べる．そして 4 節では，提案手法を用いて幾つか実験を行った．概念抽出実験，概念間の比較可能性判定実験，概念間に共通する重要な子概念の抽出実験について述べる．最後に，5 節でまとめと今後の課題について述べる．

†1 兵庫県立大学大学院工学研究科
Graduate School of Engineering, University of Hyogo

Vol.2010-DBS-151 No.33
2010/11/13

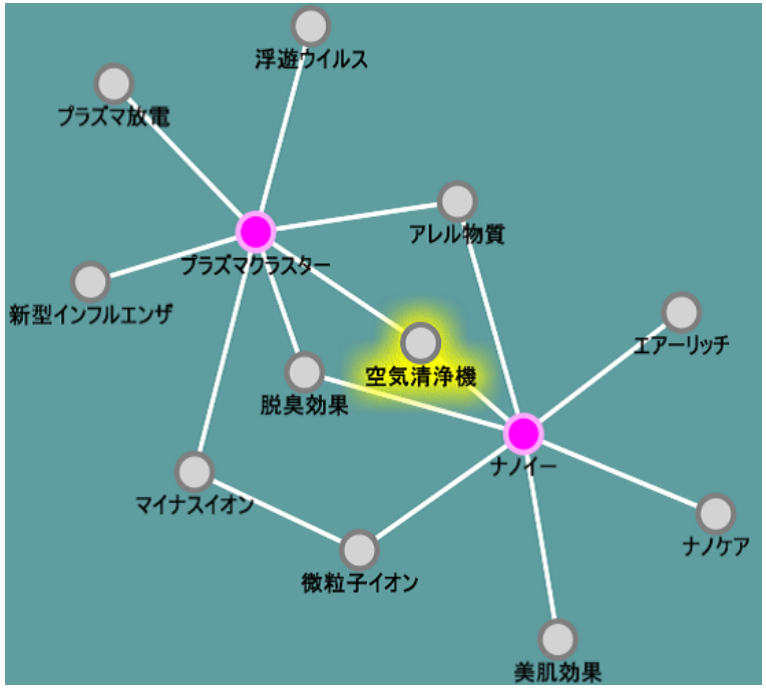


図 1 比較可能性判定のイメージ図

2. 関 連 研 究

Web サイトやテキストファイルから専門用語を抽出するためのシステムとして termex がある[1)2)]．このシステムはもともと Web サイトを対象としたメタデータ作成補助ツールとしての役割を目的としているが，それ以外にも文書から重要な語を抽出するなどの目的にも使用できる．本研究では，文書中から専門用語などの重要なキーワードを抽出するために termex を用いる．

青木らは，物事を比較する際の観点を表すキーワードである比較観点に注目し，これを抽出・提示する手法を提案している[3)]．さらに，ユーザが２つのオブジェクトを表すキーワードを与えることで，比較を行っている Web ページと比較観点を提示するシステムも提案しており，これによりユーザは２つのオブジェクトの比較を効率的かつ多角的に行うことができる．本提案手法は，１つの概念を説明している文から重要な語を発見するものであるため，発見する対象が異なるが，比較観点の発見手法と本提案手法を組合せることで，より効率的に比較判定を行えるのではないかと考えている．

山名らは，定期的に収集した blog 記事から，ユーザの比較したい対象の情報や対象に類似する物の情報を含めて検索し，提示するシステムを提案している[4)]．対象物の属性に着目し，blog 記事から属性と属性値を抽出し，これらの属性情報を用いて類似品を発見している．本提案手法は，概念を説明している文から重要な語を抽出し，これを用いて概念間の比較を行う．属性情報以外にも比較の際に重要な語が存在するため，これを用いることで，より多角的な比較判定を行うことができるのではないかと考えている．

3. 概念間の共通性に基づいた比較可能性判定手法

概念間に共通する重要な特徴が存在すれば比較ができると考え，これを比較可能性と定義する．比較可能性とは概念が持つ重要な特徴が共通しているかどうかを表すものであり，似ているかどうかを表す指標である類似性とは異なる．比較可能性の判定には図 2 のようなモデルを用いる．入力をルートの概念として扱い，概念と説明語を用いた重み付き有効グラフで表現する．エッジは関係性の強さを表しており，重要な特徴ほど重みが高くなる．ここで，概念は一般的な言葉を複数用いて説明できる語であり，その説明には他の概念や説明語を用いて表現される．説明語とは一般的に用いられる言葉のことで，それ以上の説明が不要な語である．各概念から重要な特徴を抽出し，その共通性から比較可能性の判定を行う．3.1 節でグラフモデルの生成について述べ，3.2 節で概念間の比較可能性判定手法について述べる．

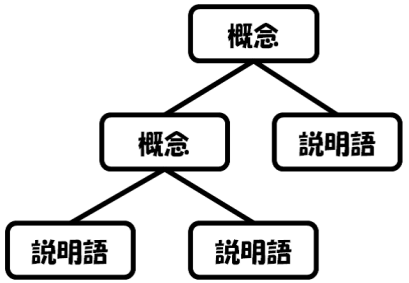


図 2 比較可能性判定のモデルの例

3.1 説明語による概念のモデル

以前，我々は説明語による概念のモデルについて提案した[6]．以下，3.1.1 節でグラフモデルの構築に必要な特徴語の抽出について述べ，3.1.2 節でグラフモデルの生成について述べる．

3.1.1 概念の特徴語抽出

クエリの入力から概念の特徴語を抽出し，概念と説明語に分類するまでの処理の流れを以下に示す．

(1) Web 上からのページの収集
入力としてクエリを 1 つ利用し，検索エンジンを用いてクエリの語を説明している文（以下，説明文とする）が出現する Web ページを収集する．まず，クエリの語 q を用いて検索を行い，検索結果として得られた Web ページ集合 $P = (p_1, p_2, \ldots, p_n)$ を収集する．

(2) 説明文の抽出
（1）で収集した Web ページ集合 P から説明文集合 $S = (s_1, s_2, \ldots, s_m)$ を抽出する．説明文 s_j は Web ページ p のテキスト内のクエリの語 q の周辺の文であり，説明文中に概念の特徴語が出現すると考えられる．クエリの語 q が出現した文からそれ以降に出現した 2 つ目の文までを説明文とする．

(3) 特徴語の決定
（2）で取得した説明文集合 S を形態素解析し，名詞集合 $Noun = (noun_1, \ldots, noun_l)$ を抽出し，一般語などの不要な語を除去したものをクエリ q の特徴語とする．

(4) 特徴語の分類
（3）で決定した特徴語を概念と説明語に分類する．具体的には，専門用語などの重要な特徴語を概念とし，それ以外の語を説明語とする．しかし，このような単純な方法では概念と判定された語の中にノイズが含まれることがあるため，3.2 節で述べる概念判定手法によりノイズと思われる概念を除去する．得られた特徴語集合を用いて 3.1.2 節で述べるグラフモデルを生成する．

3.1.2 グラフモデルの生成

3.1.1 節（4）で得られた特徴語集合を用いて重み付き有向グラフを生成する．概念 c についてのグラフモデルは (1) 式，(2) 式で定義する．

$$graph(c) = (V, E) \tag{1}$$

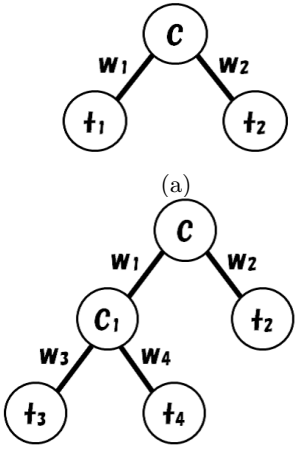
$$E = \{(t_i, t_j, weight(t_i, t_j)) | t_i, t_j \in V\} \tag{2}$$

ここで t_i，t_j は特徴語である．概念 c を親ノード，特徴語 t を子ノードとし，c-t 間の枝の重みを $weight(c,t)$ とする．ある特徴語 t_i が概念の場合は t_i を親ノードとし，同様に子ノードを追加して葉ノードが全て説明語になるまで繰り返す．例えば，概念 c の特徴語を抽出した場合，図 3 (a) のようなグラフとなる．また，特徴語 t_1 が概念であった場合，図 3 (b) のように t_1 を概念 c_1 として特徴語を抽出する．概念 c から抽出された特徴語集合を $kw(c)$ とする．また，概念 c と特徴語 t が直接つながっていない場合，重み関数 $weight(c,t)$ は c から t までの経路を表す枝集合 $path(c,t)$ を用いて (3) 式で定義する．

$$weight(c,t) = \prod_{(t_i,t_j)\in path(c,t)} weight(t_i, t_j) \tag{3}$$

$$w(c,t) = DF(c,t) \cdot IG(t) \tag{4}$$

$DF(c,t)$ は概念 c の語をクエリとして用いた検索結果の上位 N 件における特徴語 t の文書頻度である．



(a)

(b)

図 3 グラフモデル生成の例

また，一般語フィルタとして働く $IG(t)$ は（5）式で定義する．

$$IG(t) = \log\left(\frac{10^{10}}{SearchResultNumber(t) + 1}\right) \tag{5}$$

ここで，$SearchResultNumber(t)$ は特徴語 t をクエリとして用いて検索した時の検索結

果件数である．また，総文書数の代わりに十分大きな数値 10^{10} を用いた．重み $w(c,t)$ は $kw(c)$ を抽出した段階で算出し，(6) 式を用いて正規化する．このときの値が c-t 間の枝の重み $weight(c,t)$ にあたる．

$$weight(c,t) = \frac{w(c,t)}{\sum_{t' \in kw(c)} w(c,t')} \tag{6}$$

最後に，概念判定手法について述べる．概念 c から抽出された子概念 $c_1, c_2, \ldots, c_k$ 全てに対して不要な概念かどうかの判定を行う．概念 c から抽出された特徴語集合 $kw(c)$ と子概念 c_k から抽出された特徴語集合 $kw(c_k)$ を比較し，これが極端に異なる場合は重要でない概念と判断し除去する．

3.2 概念間の共通性に基づいた比較可能性

グラフモデルを用いた概念間の比較可能性の判定手法について述べる．グラフのエッジの重みが高いものほど重要な特徴であると考え，重みの高い特徴語が概念間に共通していれば比較可能であるといえる．そこで，概念と子概念間のエッジの重みが低い場合はその子概念を無視し，重みの高い子概念のみを用いて比較可能性の判定を行う．また，ルートの概念と直接繋がっている説明語は重要な語であると考え，これらを用いる．具体的には，重みの低い子概念を削除し，その個数を半分にする．例えば，図 4(a) のような概念 c から生成されたグラフを用いて説明すると，(a) の×印のように重みの低い子概念 c_2 を削除し，子概念 c_j の数を半分にする．c_2 を削除すると，図 4 (b) のようなグラフになる．次に，残った子概念 c_j の説明語 e_j と概念 c と直接つながっている説明語 e_i からベクトル v を作成する．図 4 (b) からベクトルを作成する場合，図 4 (c) のように重みを算出して説明語 e_1 と e_2 を用いてベクトルを作成する．ベクトル v の要素を説明語 e_i とし，値を c-e_i 間の枝の重み $weight(c,e_i)$ とする．このベクトルを用いて概念間のコサイン類似度を算出する．概念 c_1 についてのベクトル v_1 と概念 c_2 についてのベクトル v_2 のコサイン類似度は (7) 式により与えられる．

$$cos(c_1,c_2) = \frac{v_1 \cdot v_2}{\|v_1\|\|v_2\|} \tag{7}$$

算出されたコサイン類似度が閾値以上であれば比較可能であるとし，閾値より低ければ比較可能性が無いとする．また，概念間の比較が可能である場合は共通する重要な特徴を発見し，ユーザに提示することで概念間の比較を行いやすくする．ここで，共通する重要な特徴は概念を用いて発見する．図 5 のように，比較可能な概念間（c_a,c_b）に共通し，かつ重要な特徴を発見するために (8) 式を用いてそれぞれの子概念間（c_x,c_y）のスコアを算出する．

$$score(c_a,c_b,c_x,c_y) = weight(c_a,c_x) \times weight(c_b,c_y) \times cos(c_x,c_y) \tag{8}$$

$score$ は $[0,1]$ の値をとり，この値が高いものほど概念間に共通する重要な特徴といえる．このようなスコアにより順位付けを行い，上位 N 件を概念間に共通する重要な特徴としてユーザに提示する．各々の概念が持つ重要な特徴や概念間に共通する重要な特徴を知ることで，概念間の比較を容易に行うことができる．

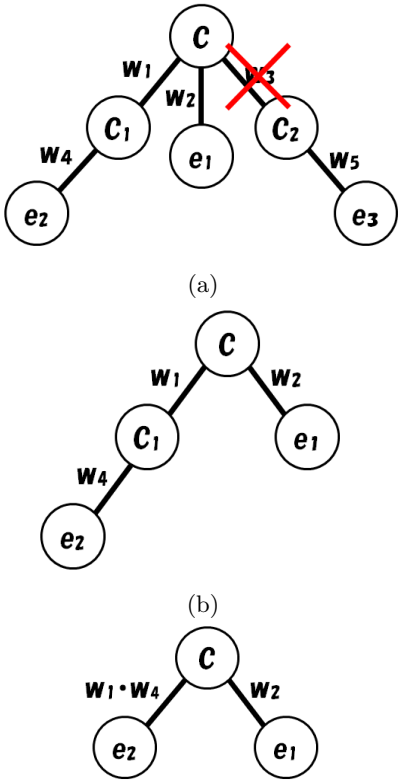


図 4 グラフモデルを用いた概念間の比較

Vol.2010-DBS-151 No.33
2010/11/13

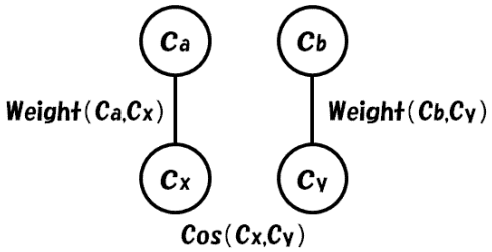


図 5 概念間に共通する重要な子概念の発見

4. 実　　験

3 節で提案した概念間の比較可能性判定手法を用いて実験を行った．4.1 節では入力となるクエリの語から子概念を抽出する実験について述べ，4.2 節では概念間の比較可能性の判定実験について述べ，4.3 節では概念間に共通する重要な子概念の抽出実験について述べる．入力として,「プラズマクラスター」,「ナノイー」,「光速ストリーマ」,「ウォーターヒート技術」,「過熱水蒸気調理」を与え，各概念間の比較可能性の判定を行う．この入力のグループを「(a) 製品の機能」とする．また，上記と同様に，入力として「吉野家」,「なか卯」,「すき家」,「餃子の王将」,「スターバックス」(以下，スタバ),「ドトール」,「マクドナルド」(以下，マクド),「モスバーガー」(以下，モス),「ロッテリア」を与え，各概念間の比較可能性の判定を行う．この入力のグループを「(b) 飲食店」とする．各入力について収集する Web ページは検索結果の上位 100 件とし，検索エンジンは Yahoo! JAPAN を用いた．

4.1 概念抽出実験

入力となるクエリの語 q から子概念を抽出する実験を行った．クエリの語 q を「プラズマクラスター」,「ナノイー」,「光速ストリーマ」,「ウォーターヒート技術」,「過熱水蒸気調理」,「吉野家」,「なか卯」,「すき家」,「餃子の王将」,「スターバックス」,「ドトール」,「マクドナルド」,「モスバーガー」,「ロッテリア」とし，検索エンジンを用いて Web ページを収集した．次に，収集した Web ページを解析し，説明文を抽出した．ここで，説明文の抽出精度を上げるためにクエリの語 q の末尾に “とは” を付加する．そして得られた説明文を形態素解析し，termex を用いて特徴語を抽出した．形態素解析には茶筌を用いた．また，処理時間短縮のため DF 値が 5 以下の特徴語はノイズとして排除し，入力された概念とその概念から抽出された概念 (子概念) について特徴語を抽出する．次に，得られた特徴語を概念と説明語に分類する．概念となる特徴語は，茶筌により「名詞-一般」,「名詞-サ変接続」,「未知語」と判定された 2 語以上からなる複合名詞とし，概念と判定されなかった特徴語を説明語とする．各入力から抽出された概念を表 1，表 2 に示す．ここで示した概念は語の重み $weight$ で降順にソートした．語の重み $weight$ は（4）式で算出する．

表 1　概念抽出実験結果
(a) 製品の機能

入力	入力から抽出された概念
プラズマクラスター	プラズマ放電，カビ菌，アレル物質，浮遊菌，空中除菌技術，水分子，空気清浄機，浮遊ウイルス，浮遊カビ菌，濃度化，酸化力，プラスイオン，イオン個数，マイナスイオン，イオン濃度，風量，脱臭効果，適用床面積，ハウスダスト，新型インフルエンザ，加湿器，空気清浄器，掃除機
ナノイー	水分量，微粒子イオン，アレル物質，マイナスイオン，空気清浄機，弱酸性，体積比，カビ菌，長寿命，発生ユニット，ナノケア，エアーリッチ，加湿空気清浄機，加湿器，脱臭効果，加湿機能，加湿セラミックファンヒーター，空気イオン，風量，抑制効果，発生装置，美肌効果，空気清浄
光速ストリーマ	酸化分解力，高速電子，プラズマ放電，カビ菌，アレル物質，放電領域，放電技術，放電量，光クリエール，酸化分解速度，空気清浄機，光速技術，太陽表面温度，ウイルス除去，加湿機能，空気環境，ニオイ成分，加湿空気清浄機，光触媒
ウォーターヒート技術	過熱水蒸気，減塩，酸素調理，ヘルシオ加熱，オーブン加熱，ウォーターオーブン，レンジ加熱，自動調理，カロリーダウンメニュー，カラー液晶，過熱水蒸気調理，自動メニュー
過熱水蒸気調理	過熱水蒸気，過熱水蒸気調理機能，ウォーターヒート技術，減塩，電子レンジ

表 1，表 2 の実験結果より，入力された語に関連のある概念が幾つか抽出されていることが分かる．しかし,「号店」や「発売開始」など，概念として不適切な語も幾つか含まれているので，これらの語を除去する方法を考える必要がある．

4.2 比較可能性判定実験

4.1 節で抽出された概念と説明語を用いて，比較可能性判定手法により各入力語間の比較可能性の判定を行った．提案手法を用いて各入力語間のコサイン類似度を算出した結果を表 3 に示す．概念間のコサイン類似度が閾値以上の場合を比較可能であるとし，閾値より低ければ比較可能ではないとする．人手で正解例を作成し，提案手法についての実験結果から適合率，再現率を算出した．人手で作成した正解例を表 5 に示す．概念間の比較が可能であると判断した場合は○，比較ができないと判断した場合は×とした．また，適合率，再現率を

Vol.2010-DBS-151 No.33
2010/11/13

表 2 概念抽出実験結果
(b) 飲食店

入力	入力から抽出された概念
吉野家	牛丼，牛丼屋，豚丼，すき家，牛鮭定食，なか卯，並盛り，牛丼チェーン，ショートプレート，字テーブル，牛皿，並盛，牛丼並，鶏丼，店舗数，同業他社，牛めし，特盛，牛鍋丼，号店，すき屋，チェーン店
すき家	牛丼，なか卯，牛丼チェーン，牛丼屋，並盛，株式会社ゼンショー，豚丼，是正申告，牛丼並盛り，メガ牛丼，牛丼店，牛丼チェーン店，牛丼大手，すき屋，大手牛丼チェーン，牛丼業界，輸入再開，牛丼戦争，商品用，チーズ牛丼，牛皿，特盛，並盛り，チェーン店，牛めし，店舗数
なか卯	牛丼，親子丼，すき家，牛丼屋，和風牛丼，丼ぶり，すき屋，チェーン店，牛丼チェーン，うどん屋
餃子の王将	中華料理
スタバ	スターバックスコーヒー，スターバックスコーヒージャパン，スターバックス社，コーヒー豆，号店，チェーン店，コーヒーショップ，スターバックスカード，当地タンブラー，店舗数，スターバックス店舗，カフェモカ，コーヒー店
ドトール	レスホールディングス，コーヒー豆，スターバックス，コーヒーチェーン，コーヒー農園，株式会社ドトールコーヒー，アイスコーヒー，ブレンドコーヒー，コーヒーショップ，店舗数，号店
マクド	ハンバーガーチェーン，フランチャイズ権，外食産業，ファーストフード，ハンバーガーショップ，消費者，ドライブスルー，チェーン店，号店，ハンバーガー大学，従業員，効率性，ファーストフード業界，グローバル化，店舗数，合理性，品質管理，世界大戦，先進国，管理職
モス	ライスバーガー，ミートソース，ハンバーグサンド，モスド，ファーストフード店，ラー油バーガー，ミスタードーナツ，ファーストフード，号店
ロッテリア	絶品チーズバーガー，エビバーガー，チーズバーガー，株式会社ロッテリア，店舗限定，モスバーガー，ファーストフード，えびバーガー，発売開始

表 5 に示す．ここで，適合率，再現率は (9) 式，(10) 式を用いて算出した．また，提案手法におけるコサイン類似度の閾値を 0.2 と設定した．

$$\text{適合率} = \frac{\text{人手で作成した正解例と一致した数}}{\text{全組合せ数}} \tag{9}$$

$$\text{再現率} = \frac{\text{人手で作成した正解例の○と一致した数}}{\text{人手で作成した正解例の○の数}} \tag{10}$$

空気清浄機能である「プラズマクラスター」,「ナノイー」,「光速ストリーマ」間のコサイン類似度は高いと予想していたが，提案手法を用いた結果 (表 3) より，低いことが分かる．これは，空気清浄機能以外のセールスポイントがそれぞれ異なっていたからではないかと考

表 3 提案手法による各概念間のコサイン類似度
(a) 製品の機能

	1	2	3	4
1．プラズマクラスター				
2．ナノイー	0.33			
3．光速ストリーマ	0.21	0.21		
4．ウォーターヒート技術	0.02	0.05	0.01	
5．過熱水蒸気調理	0.00	0.01	0.00	0.60

(b) 飲食店

	1	2	3	4	5	6	7	8
1．吉野家								
2．なか卯	0.26							
3．すき家	0.48	0.61						
4．餃子の王将	0.02	0.01	0.01					
5．スタバ	0.11	0.04	0.06	0.01				
6．ドトール	0.07	0.03	0.05	0.01	0.31			
7．マクド	0.14	0.03	0.07	0.01	0.13	0.06		
8．モス	0.12	0.04	0.06	0.02	0.08	0.05	0.24	
9．ロッテリア	0.12	0.05	0.06	0.01	0.09	0.05	0.33	0.61

表 4 人手で作成した正解例
(a) 製品の機能

	1	2	3	4
1．プラズマクラスター				
2．ナノイー	○			
3．光速ストリーマ	○	○		
4．ウォーターヒート技術	×	×	×	
5．過熱水蒸気調理	×	×	×	○

(b) 飲食店

	1	2	3	4	5	6	7	8
1．吉野家								
2．なか卯	○							
3．すき家	○	○						
4．餃子の王将	×	×	×					
5．スタバ	×	×	×	×				
6．ドトール	×	×	×	×	○			
7．マクド	×	×	×	×	○	○		
8．モス	×	×	×	×	×	×	○	
9．ロッテリア	×	×	×	×	○	○	○	○

表 5 比較可能性判定手法の精度

	適合率	再現率
(a) 製品の機能	1.00(10/10)	1.00(4/ 4)
(b) 飲食店	0.89(32/36)	0.64(7/11)
(a)+(b) 全て	0.91(42/46)	0.73(11/15)

えられる．その結果，全体的にコサイン類似度が低くなっている．過熱水蒸気を用いた調理機能である「ウォーターヒート技術」,「過熱水蒸気調理」間のコサイン類似度は重要な概念が共通しているため高いことが分かる．牛丼，豚丼チェーンストアとして知られている「吉野家」,「なか卯」,「すき家」間のコサイン類似度は高いと予想していたが，提案手法を用いた結果 (表 4) より,「吉野家」,「なか卯」間のコサイン類似度は低い．その理由として次のことが考えられる．各概念のみが持つ独自の特徴語（例えば，なか卯では「うどん」，すき家では「カレー」関連の語やメニュー名）の影響が大きく，コサイン類似度が低かったのではないかと考えられる．また，コーヒーショップとして知られている「スターバックス」,「ドトール」間のコサイン類似度が低い．これは,「ドトール」から抽出された説明語にはコーヒー関連の語が多かったが,「スターバックス」から抽出された説明語にはメニュー名などの「スターバックス」のみで扱っているコーヒー関連の語が多かったためであると考えられる．ファーストフード店として知られている「マクドナルド」,「モスバーガー」,「ロッテリア」については，メニュー名などの説明語が抽出されていたが，その材料となる語も説明語として抽出されていたため，コサイン類似度が比較的高かった．表 5 の適合率，再現率から，人手により作成した正解例と実験結果が概ね一致していることが分かる．しかし,「マクドナルド」,「ロッテリア」からコーヒーに関連する特徴語が抽出されると予想していたが，少なかったため,「スターバックス」や「ドトール」とのコサイン類似度が低下し，正解例と一致しなかった．これは「マクドナルド」,「ロッテリア」にとって，コーヒーよりもハンバーガーに関連する特徴語の方が重要であったからだと思われる．

4.3 概念間に共通する重要な子概念の抽出実験

概念間に共通する重要な特徴を抽出する実験を行った．以下に示す比較可能であると判定された 7 組の概念について実験を行うものとする．

- プラズマクラスター，ナノイー
- すき家，吉野家
- すき家，なか卯
- なか卯，吉野家
- ウォーターヒート技術，過熱水蒸気調理
- スターバックス，ドトール
- iPhone，Xperia

3.2 節で述べた提案手法を用いて *score* を算出し，その上位 5 件を人手により評価した．評価は 6 人で行う．5 段階で評価を行い，適切であると判断した場合は 5，不適切であると判断した場合は 1，どちらともいえない場合は 3 とする．適切（5 または 4）と評価された子概念を重要な特徴と考え,(11) 式を用いて精度を算出する．

$$\text{精度} = \frac{\text{適切（5 または 4）と評価された子概念の数}}{\text{評価された子概念の総数}} \tag{11}$$

実験結果から求めた精度および評価者全員が適切と評価した子概念を表 6 に示す．

表 6 共通する重要な特徴の評価

	精度	評価者全員が適切と評価した子概念
プラズマクラスター，ナノイー	0.67(20/30)	マイナスイオン，空気清浄機
すき家，吉野家	0.70(21/30)	牛丼，牛皿
すき家，なか卯	0.57(17/30)	牛丼，牛めし
なか卯，吉野家	0.60(18/30)	牛丼
ウォーターヒート技術，過熱水蒸気調理	0.57(17/30)	過熱水蒸気，減塩
スターバックス，ドトール	0.50(15/30)	コーヒー豆
iPhone，Xperia	0.70(21/30)	タッチパネル，タッチスクリーン
全体	0.61(129/210)	

実験結果（表 6）より，約 6 割が概念間に共通する重要な特徴と評価されていることがわかる．多くの評価者から不適切（1 または 2）と評価された子概念は，4.1 節の実験結果の段階で抽出された不適切な概念であったので，概念判定アルゴリズムを改良することで改善できると考えている．また，今回の実験で扱った全ての概念間において，評価者全員から適切（5 または 4）と評価された子概念が最低 1 つ存在しており，比較可能な概念間において，共通する重要な特徴が抽出できていることがわかる．

5. まとめと今後の課題

本論文では，Web 上での概念間の共通性に基づいた比較可能性の判定を行う手法を提案した．比較可能性とは概念が持つ重要な特徴が共通しているかどうかを表すものであり，似ているかどうかを表す指標である類似性とは異なる．概念は一般的な言葉を複数用いて説明できる語であり，その説明には他の概念や説明語を用いて表現される．説明語とは一般的に

Vol.2010-DBS-151 No.33
2010/11/13

用いられる言葉のことで，それ以上の説明が不要な語である．本研究で扱う概念は商品や機能などに限定している．概念と説明語を用いて重み付き有向グラフを生成し，概念間の比較可能性の判定を行っている．比較可能性の判定には概念と説明語からなる重み付き有効グラフを用いる．また，3 節で提案した手法を用いて，概念抽出実験，概念間の比較可能性判定実験，概念間に共通する重要な子概念の抽出実験を行い，評価を行った．抽出された概念の中には，直接関係の無い語が含まれていたため，概念を決定するアルゴリズムを改良する必要があると考えられる．概念間に共通する重要な子概念の抽出実験により，比較可能な概念間において，共通する重要な特徴が抽出できていることがわかった．今後は，ユーザへの提示方法などを検討し，概念間を比較しやすいユーザインタフェースの開発を行うことを考えている．しかし，現在は処理時間が長いため，結果が出力されるまでに時間が掛かってしまう．これは抽出される概念が多ければ多いほど処理が多くなるからである．出来る限り精度を損なわないようにし，処理時間を短くする必要がある．

謝　　辞

本研究の一部は，平成 22 年度科研費基盤研究 (B)(2)「ユーザの潜在的意図を用いたレス・コンシャス情報検索基盤の構築」(課題番号：20300039) によるものです．ここに記して謝意を表すものとします．

参 考 文 献

1) 小島博之，前田朗．キーワード（専門用語）自動抽出システムの構想とその展開．第 51 回日本図書館情報学会研究発表要綱．pp.17-20，2003.
2) `http://gensen.dl.itc.u-tokyo.ac.jp/win.html`.
3) 青木伸也，湯本高行，新居学，高橋豊．Web 上の比較表現を用いた 2 オブジェクト間の比較観点の発見，DEWS2008 A7-6.
4) 山名健悟，西村圭亮，滝沢敏裕，湯浅将英，大山実．blog 検索と類似品情報を用いた選定支援システム．2005-DD-52(3)．pp.17-21(2005).
5) D. Bollegala, Y. Matsuo, and M. Ishizuka.Measuring semantic similarity between words using web search engines. *In Proceedings of the 16th International World Wide Web Conference (WWW-07)*, pp.757-766, 2007
6) 岡田崇，湯本高行，新居学，高橋豊，Web 上の周辺語句を用いた概念間の比較可能性の判定 ，DEIM フォーラム 2010，C1-4，(2010).