

Web 上の周辺語句を用いた概念間の比較可能性の判定

岡田 崇[†] 湯本 高行[†] 新居 学[†] 高橋 豊[†]

[†] 兵庫県立大学院工学研究科

〒 671-2280 兵庫県姫路市書写 2167

E-mail: [†]er09q010@steng.u-hyogo.ac.jp, ^{††}{yumoto,nii,takahasi}@eng.u-hyogo.ac.jp

あらまし 本論文では Web 上の周辺語句を用いて、概念間の比較可能性の判定を行う手法を提案する。概念は一般的な言葉を複数用いて説明できる語であり、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。また、本研究で扱う概念は商品や機能などに限定している。比較可能性とは概念間に共通する特徴が存在するかどうかを表すものである。概念間の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向グラフを用いる。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。

キーワード Web 上の周辺語句, 比較可能性, 概念

Judging Comparability between Concepts using Surrounding Text on the Web

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

[†] Graduate School of Engineering, University of Hyogo

2167 Syosya, Himeji, Hyogo 671-2280, Japan

E-mail: [†]er09q010@steng.u-hyogo.ac.jp, ^{††}{yumoto,nii,takahasi}@eng.u-hyogo.ac.jp

Abstract In this paper, we propose a method to judge comparability between concepts using surrounding text on the Web. The concept is term a explained with general terms. The concept represents by other concepts and explanation terms. Explanation terms are general terms which do not need any more explain about concept. Our targets of concepts are products and function in this paper. Comparability means a metric of common features between concepts. The method to judge comparability between concepts uses weighted digraph consisting of concepts and explanation terms. We extract keywords from sentences to explain features of concepts and judge comparability by sets.

Key words Surrounding Text on the Web, Comparability, Concepts

1. はじめに

近年、商品やサービスを購入するためにインターネットを利用して情報を収集する人が増加している。この原因としてインターネット上で商品を購入することができるショッピングサイトの増加やネット普及率の増加などが考えられる。ネットショッピングには時間・場所を問わずに商品を購入できるといったメリットがあり、利用者数も年々増加傾向にある。インターネットを利用して商品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あげられる。検索エンジンではインターネットで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け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。なぜならユーザ自身が知らない商品については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を提示することで、ユーザへの負担を軽減することができると考えている。本稿では、概念間の比較が可能かどうか判定する手法を提案する。概念は一般的な言葉を複数用いて説明できる語であり、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おり、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。

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、概念を抽出する実験および概念間の比較可能性を判定する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。

2. 関連研究

小山らは、Web ページの文書構造を利用し、与えられた語を詳細化するような語を発見する手法を提案している [1]。タイトルと本文という Web ページの文書構造に着目し、ある語がタイトルとして用いられている文書の本文に詳細語が現れると仮定することで、与えられた語を詳細化するキーワードを精度良く抽出することができる。キーワード A, B が Web ページに出現する位置のパターンを図 1 に示す。

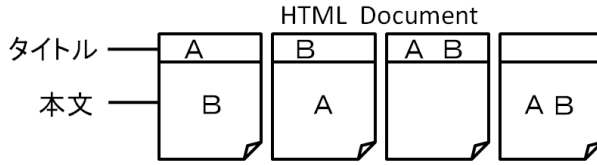


図 1 Web ページの文書構造

以下の条件を満たすとき、キーワード A はキーワード B を詳細化するキーワードである。

$$\frac{DF(intitle(A) \wedge B)}{DF(intitle(A))} > \frac{DF(A \wedge B)}{DF(A)}$$

ここで、 $DF(intitle(A))$ はタイトルにキーワード A を含むページの数、 $DF(intitle(A) \wedge B)$ はタイトルにキーワード A を含み、かつページのどこかにキーワード B を含むページの数である。また、偶然によるものなのか意味のあるものなのかを測定するため、以下のような χ^2 検定を行う。

$$\chi_0^2 = \sum_{i=1}^2 \sum_{j=1}^2 \frac{x_{ij} - a_i b_j / N}{a_i b_j / N}$$

なお、変数の定義は表 1 に示す。

表 1 変数の定義

$N = DF(A)$
$a_1 = DF(intitle(A))$
$a_2 = N - a_1$
$b_1 = DF(A \wedge B)$
$b_2 = N - b_1$
$x_{11} = DF(intitle(A) \wedge B)$
$x_{12} = a_1 - x_{11}$
$x_{21} = b_1 - x_{11}$
$x_{22} = N - a_1 - b_1 + x_{11}$

χ_0^2 の値は自由度 1 の χ^2 分布に従う。ここで、有意水準を α に設定したとすると、 $\chi_0^2 < \alpha$ であれば、2 つのキーワード A, B の間に関連があるということがいえる。本研究では、入力された語と説明文から抽出された単語との関係を判定するためにこの手法を用いている。

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 [2] [3]。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 [4]。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み合わせることで、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 [5]。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. 概念間の比較可能性判定手法

本研究では、入力としてクエリを 1 つ利用し、それと比較可能な概念の発見を目的とする。比較可能性の判定には図 3 のようなモデルを用いる。入力をルートの概念として扱う。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。概念と説明語の定義およびグラフモデルの生成については 3.1 節で述べる。

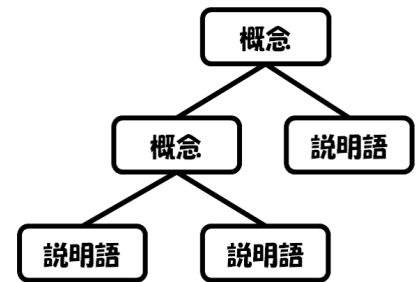


図 3 比較可能性の判定のモデルの例

3.1 節では、概念を比較する際に必要となる語を抽出する手法およびグラフモデルの生成について述べる。3.2 節では、概念間の比較が可能かどうか判定する手法について述べる。

3.1 説明語による概念のモデル

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決定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．

(2) 説明文の抽出

(1) で収集した Web ページから説明文を抽出する．説明文中に概念の特徴語が出現すると考えられる．

(3) 特徴語の決定

(2) で取得した説明文から名詞を抽出し，一般語などの不要な語を除去し，特徴語を決定する．

(1) Web 上からのページの収集

検索エンジンを用いて，概念の性能や特徴を説明している文が書かれたページを収集する．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 $P = (p_1, p_2, \dots, p_n)$ を収集する．

(2) 説明文の抽出

(1) で収集した Web ページ集合 $P = (p_1, p_2, \dots, p_n)$ から概念の性能や特徴を説明している文の集合 $S = (s_1, s_2, \dots, s_m)$ 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) で収集した説明文集合 $S = (s_1, s_2, \dots, s_m)$ を形態素解析し，名詞集合 $Noun = (noun_1, noun_2, \dots, noun_i)$ を抽出する．次に，得られた名詞 $noun$ がクエリの語 q に関係のある語か判定するために“話題を詳細化するキーワードを発見するアルゴリズム” [1] を用いる．全ての名詞 $noun$ に対して判定を行い，クエリの語 q の詳細語だと判定された名詞 $noun$ を特徴語とする．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) で抽出した特徴語を用いて重み付き有向グラフを生成する．まず，特徴語を概念と説明語に分ける．概念はある分野でのみ使用される語であり，説明が必要な語である．概念の説明には，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．具体的には，複合名詞などが概念であり，それ以外の語を説明語とする．概念 c についてのグラフモデルは (1) 式，(2) 式で定義する．

$$graph(c) = (V, E) \quad (1)$$

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

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 4 (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 4 (b) のように t_1 を概念 c_1 として特徴語を抽出する．

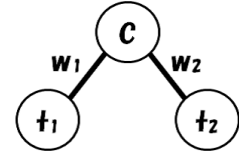
概念 c から抽出された特徴語集合を $kw(c)$ とする．また，概念 c と特徴語 t が直接つながっていない場合，重み関数 $weight(c, t)$ は c から t までの経路を表す枝集合 $path(c, t)$ を用いて (3) 式で定義する．

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

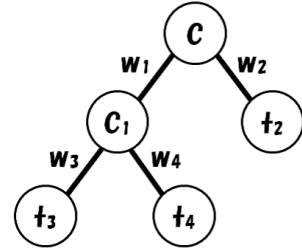
例えば，図 5 のような重み付きグラフが生成された場合， $c-e_1$ 間の枝の重みは $weight(c, e_1) = w_2$ である．また， c から e_2 までの経路を表す枝集合は $path(c, e_2) = path(c, c_1) + path(c_1, e_2)$ であり， $c-e_2$ 間の枝の重みは $weight(c, e_2) = weight(c, c_1) \times weight(c_1, e_2) = w_1 \times w_3$ である．概念 c における特徴語 t の重み $w(c, t)$ は (4) 式を用いて算出する．

$$w(c, t) = DF(c, t) \cdot IG(t) \quad (4)$$

$DF(c, t)$ 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 4 グラフモデル生成の例

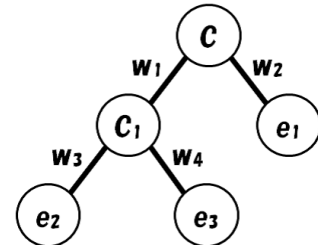


図 5 重み付きグラフ

また，一般語フィルタとして働く $IG(t)$ は (5) 式で定義する．

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

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．

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

例えば，図 5 の重み付きグラフを生成する場合， c_1 と e_1 が抽出された段階で (4) 式を用いて w_1 と w_2 を算出し，(6) 式を用いて正規化する．このとき，(6) 式の分母は $w_1 \times w_2$ となる．同様に，次は e_2 と e_3 を抽出し， w_3 と w_4 を算出し，正規化する．

3.2 概念間の比較可能性

3.1 節で生成したグラフモデルを用いて概念間の比較可能性の判定を行う．比較可能性とは概念間に共通する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となる類似とは異なる．

3.2.1 セールスポイントを重視した比較

セールスポイントとは，概念が持つ重要な特徴や性能を表す語のことである．つまり，概念の特徴語の中で重みの高い語のことであり，これらの語が類似している概念は比較可能性があると考えられる．そこで，セールスポイントを重視した比較では， c - c_j 間の枝の重み $weight(c, c_j)$ が低い場合は概念 c_j を無視し，重みの高い場合は概念 c_j を用いて比較可能性の判定を行う．また，比較対象となる概念 c と直接つながっている（つまり， $weight(c, e_i)$ が存在する）説明語 e_i は重要な語であると考え，これらを用いる．従って， $kw(c)$ における重みの高い概念 c_j と説明語 e_i を用いて比較可能性の判定を行う．まず，重み $weight(c, c_j)$ の低い概念 c_j を削除し， c_j の個数を半分ににする．例えば，図 6 (a) のように概念 c_2 を削除し，概念 c_j の数を半分ににする． c_2 を削除すると，図 6 (b) のようなグラフになる．次に，残った概念 c_j の説明語 e_j と $kw(c)$ における説明語 e_i からベクトル v を作成する．図 6 (b) からベクトルを作成する場合，図 6 (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．

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

3.2.2 バランスを重視した比較

概念 c についてグラフモデルを生成したとき，その末端の子ノードは概念 c を最も一般的な語で表現した語の集合であると考えられる．これらの語が類似している概念は比較可能性があると考えられる．そこで，バランスを重視した比較では，概念 c_j は用いずに末端の子ノード，つまり説明語 e_i のみを用いて比較可能性の判定を行う．例えば，図 6 (b) のようなグラフが生成された場合，末端の子ノードのみを用いるため，図 6 (c) のようなグラフとなる．また，全ての重み $weight(c, e_i)$ を 1

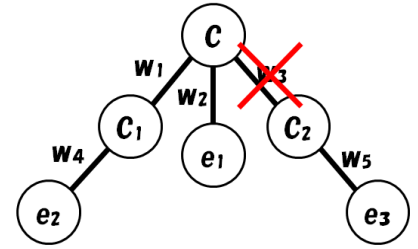
として概念間の Jaccard 係数を算出する．概念 c_i と概念 c_j の Jaccard 係数は (8) 式で定義する．

$$Jaccard(explain(c_i), explain(c_j)) = \frac{|explain(c_i) \cap explain(c_j)|}{|explain(c_i) \cup explain(c_j)|} \quad (8)$$

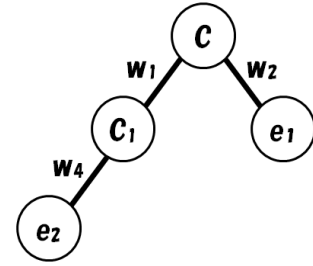
なお，概念 c についての説明語集合 $explain(c)$ は (9) 式で定義する．

$$explain(c) = \{e_i | e_i \in kw^*(c)\} \quad (9)$$

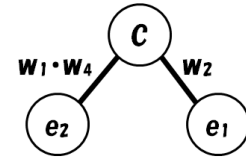
ここで， $kw^*(c)$ は概念 c および概念 c_j から抽出される全ての説明語集合とする．



(a)



(b)



(c)

図 6 グラフモデルを用いた概念間の比較

4. 実 験

3. 節で提案した手法を用いて，実験を行った．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

クエリ 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。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」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 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 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。入力を飲食店名としたため、メニュー名や商品名が抽出されている。また、抽出される概念や説明語の数にばらつきがあるため、比較可能性の判定に問題が出るのではないかと考えられる。

表 2 概念抽出実験結果

入力	入力から抽出された概念
プラズマクラスター	カビ菌, 水分子, 濃度化, 風量, 中央付近, イオン個数, プラズマ放電, イオン濃度, 濃度イオン, 長寿命, ハウスダスト, 床面積, マイナスイオン
ナノイー	カビ菌, 水分量, マイナスイオン, 弱酸性, ナノサイズ
光速ストリーマ	高速電子, カビ菌, プラズマ放電, 大風量, 光クリエール, 光速ストリーマユニット, 中間期, 空気環境, 光触媒, ウイルス対策, 熱エネルギー, 高速バス
ウォーターヒート技術	減塩, 高温状態, 酸素状態, 食品内部, カラー液晶, 自動メニュー, 中心部
過熱水蒸気調理	スチーム調理, 赤効果, 電子レンジ, スチームオープン, 角皿, 食品加工, 熱エネルギー
吉野家	牛丼, 豚丼
すき家	牛丼, 豚丼, なか卯, 牛丼チェーン, 牛丼屋, 店舗数, すき屋
なか卯	親子丼, 牛丼
餃子の王将	チェーン店
スタバ	スターバックスコーヒー, コーヒー豆, スターバックスカード, スターバックスコーヒージャパン
ドトール	コーヒー豆
マクド	メガマック, バリューセット, 期間限定
モス	モスチーズバーガー, 緑モス, ファーストフード, 期間限定
ロッテリア	エビバーガー, 絶品チーズバーガー

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、各入力語間の Jaccard 係数とコサイン類似度の算出を行った。提案手法を用いて各入力語間の Jaccard 係数を算出した結果を表 3 に、コサイン類似度を算出した結果を表 4 に示す。また、比較対象として、得られた特徴語について $WebPMI$ [6] を算出し、これを用いて概念間の近さを求めた。概念 p, q 間の $WebPMI(p, q)$ は (10) 式で定義される。

$$WebPMI(p, q) = \log_2 \left(\frac{\frac{H(p \cap q)}{N}}{\frac{H(p)}{N} \frac{H(q)}{N}} \right) \quad (10)$$

ここで、 $H(p)$ は p をクエリとしたときの検索結果件数であり、 N は Web 全体のページ数である。今回は $N = 10^{10}$ と設定した。WebPMI を用いて各入力語間の近さを算出した結果を表 5 に示す。

概念間の Jaccard 係数、コサイン類似度が閾値を越えた場合に比較可能性があるとし、閾値より低ければ比較可能性が無いとする。人手で正解例を作成し、提案手法及び WebPMI 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 6 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、再現率を表 7 に示す。ここで、適合率、再現率は (11) 式、(12) 式を用いて算出する。

$$\text{適合率} = \frac{\text{人手で作成した正解例と一致した数}}{\text{全組合せ数}} \quad (11)$$

$$\text{再現率} = \frac{\text{人手で作成した正解例の○と一致した数}}{\text{人手で作成した正解例の○の数}} \quad (12)$$

また、提案手法及び WebPMI における閾値はそれぞれ異なる値を設定した。提案手法における Jaccard 係数の閾値を 0.3、コサイン類似度の閾値を 0.2 と設定した。WebPMI における概念間の近さの閾値を 7.0 と設定した。

表 3 提案手法による各概念間の Jaccard 係数

(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	0.31			
3. 光速ストリーマ	0.16	0.14		
4. ウォーターヒート技術	0.03	0.03	0.02	
5. 過熱水蒸気調理	0.00	0.00	0.00	0.09

(b) 飲食店

	1	2	3	4	5	6	7	8
1. 吉野家								
2. なか卯	0.41							
3. すき家	0.60	0.58						
4. 餃子の王将	0.03	0.08	0.04					
5. スタバ	0.02	0.03	0.03	0.02				
6. ドトール	0.02	0.06	0.05	0.06	0.46			
7. マクド	0.00	0.06	0.07	0.04	0.03	0.05		
8. モス	0.04	0.06	0.05	0.09	0.02	0.05	0.12	
9. ロッテリア	0.00	0.05	0.04	0.04	0.00	0.03	0.10	0.16

表 4 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	0.26			
3. 光速ストリーマ	0.23	0.26		
4. ウォーターヒート技術	0.00	0.01	0.01	
5. 過熱水蒸気調理	0.01	0.01	0.00	0.49

(b) 飲食店

	1	2	3	4	5	6	7	8
1. 吉野家								
2. なか卯	0.02							
3. すき家	0.04	0.31						
4. 餃子の王将	0.04	0.02	0.07					
5. スタバ	0.06	0.03	0.05	0.06				
6. ドトール	0.03	0.01	0.04	0.04	0.24			
7. マクド	0.00	0.02	0.07	0.02	0.00	0.01		
8. モス	0.04	0.03	0.08	0.09	0.07	0.04	0.36	
9. ロッテリア	0.00	0.01	0.06	0.04	0.00	0.01	0.49	0.50

表 5 WebPMI を用いた各概念間の近さ

(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	8.94			
3. 光速ストリーマ	6.40	8.87		
4. ウォーターヒート技術	6.96	4.33	3.15	
5. 過熱水蒸気調理	2.07	1.47	0.00	9.34

(b) 飲食店

	1	2	3	4	5	6	7	8
1. 吉野家								
2. なか卯	7.11							
3. すき家	7.07	7.69						
4. 餃子の王将	5.31	6.86	5.53					
5. スタバ	4.83	6.15	4.72	4.94				
6. ドトール	5.37	7.08	5.09	5.38	5.35			
7. マクド	4.96	5.92	4.75	5.00	4.69	4.61		
8. モス	5.49	7.10	5.32	5.95	5.43	5.61	6.18	
9. ロッテリア	5.98	7.55	5.69	6.34	5.63	5.96	6.17	7.34

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 Jaccard 係数、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 (a), 表 4 (a)) より低いことが分かる。その理由として、抽出された概念の数にばらつきがあったことが挙げられる。その結果、全体的に Jaccard 係数とコサイン類似度が低く、閾値を下回ったため適合率、再現率が低くなっている。

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 Jaccard 係数、コサイン類似度は高いと予想していたが、セールスポイント重視の比較手法を用いた結果 (表 3 (b)), コサイン類似度が低い。その理由として次のことが考えられる。「吉野家」、「なか卯」、「すき家」から得ら

表 6 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 7 精 度

(a) 製品の機能		
	適合率	再現率
バランスを重視した比較	0.70(7/10)	0.25(1/4)
セールスポイントを重視した比較	1.00(10/10)	1.00(4/4)
WebPMI	0.90(9/10)	0.75(3/4)

(b) 飲食店

	適合率	再現率
バランスを重視した比較	0.92(33/36)	0.36(4/11)
セールスポイントを重視した比較	0.86(31/36)	0.45(5/11)
WebPMI	0.72(26/36)	0.36(4/11)

れる特徴語の数が 50 以下と少なかったため、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 Jaccard 係数、コサイン類似度が低かった。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、その材料となる語が説明語として抽出されていたため、Jaccard 係数やコサイン類似度が比較的高かった。

5. まとめと今後の課題

本論文では、Web 上の周辺語句を用いて、概念間の比較可能性の判定を行う手法を提案した。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．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う．比較可能性とは概念間に共通する特徴が存在するかどうかを表す指標であり，概念の本質的な特徴を扱うため，類似とは異なる．また，3. 節で提案した手法を用いて，概念を抽出する実験および概念間の比較可能性を判定する実験を行った．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．また，抽出された概念と説明語の数にばらつきがあるため，比較可能性の判定に問題が出たのではないかと考えられる．今後は，抽出する概念と説明語の数を決定し，概念間の比較可能性を判定する実験を行う必要がある．

謝 辞

本研究の一部は，平成 21 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．

文 献

- [1] Oyama, S. and Tanaka, K.: Query Modification by Discovering Topics from Web Page Structures, *Proc. 6th Asia Pacific Web Conference (APWEB'04)*, Lecture Notes in Computer Science, Vol.3007, pp.553-564(2004).
- [2] 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．
- [3] <http://gensen.dl.itc.u-tokyo.ac.jp/win.html>．
- [4] 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．
- [5] 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)．
- [6] 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)*, pages 757-766, 2007