4F1-OS-5-7

メニュー階層分かりやすさ評価のための語対の連想関係，関連関係自動判定

Automatic classification of word pairs into those associative and those related to evaluation of menu hierarchies

高尾　美代子[*1]　酒井　浩之[*2]　鶴田　雅信[*2]　増山　繁[*2]　渡邊　知早[*3]
Miyoko Takao　Hiroyuki Sakai　Masanobu Tsuruta　Shigeru Masuyama　Chihaya Watanabe
梅村　祥之[*4]
Yoshiyuki Umemura

[*1]豊橋技術科学大学　大学院　工学研究科　情報・知能工学専攻
Computer science and engineering,Toyohashi University of Technology
[*2]豊橋技術科学大学
Toyohashi University of Technology
[*3]トヨタ自動車株式会社
Toyota Motor Corporation
[*4]広島工業大学
Hiroshima Institute Technology

This paper describes automatic classification of word pairs into those associative and those related to evaluation of menu hierarchies. We prove that rank correlation coefficient is 0.7 by using mutual information between words, and Support Vector Machine.

1. はじめに

Web ページや車のナビ，携帯電話のメニューなど，世の中のあらゆるところにメニュー階層構造が用いられており，我々は日常的にメニュー階層構造の操作を行っている．操作の分かりやすいメニュー階層構造を開発するためには，メニュー階層全体での繋がりの分かりやすさを定量的に客観評価できる指標が必要である．そこで，本研究では，メニュー階層構造の分かりやすさの評価に用いるため，2 単語間の連想関係，関連関係の自動判定，および，連想度，関連度の指標化を検討した．

ここで，連想関係とは，全体-部分関係，上位-下位概念，装置-機能を含む親子関係である．関連関係とは，ある全体に対する部分集合，ある上位概念に対する下位概念の集合，ある装置に対する機能の集合を含む兄弟関係である．

また，連想度とは，ある 2 単語に関して，心理的に連想関係があると感じる度合いである．関連度は，ある 2 単語に関して，心理的に関連関係があると感じる度合いである．

2. 本研究の位置づけについて

岡本ら [1] は，被験者に刺激語（小学校学習基本語彙中の名詞から 100 語）と 7 つの課題（上位概念，下位概念，部分・材料，属性，類義語，動作，環境）を与え，連想語を収集する心理実験を行い，連想度を「単語間の距離」として定式化している．彼らの手法では，連想度を算出するためには心理実験が必要になる．それに対して本手法においては，2 単語間の連想度・関連度は，Web から取得したデータを用いた統計的手法と，Support Vector Machine(以下，SVM) による機械学習を用いて自動的に指標を推定可能である．これにより，心理実験に必要な人手のコストや時間を短縮することができる．

また，北島ら [2] は，階層メニューから項目を選択する過程を Markov 連鎖として，メニュー階層を使用するユーザのモデルを作成し，モデルを用いたインターフェースの初期の段階に適応できる階層メニュー評価法を提案している．それに対して本手法では，単語間の統計量を使用して，単語間の連想度・関連度を指標化することにより，メニュー階層評価への応用を検討する．

提案手法の概要図を図 1 に示す．提案手法の詳細については 3.3 節で述べる．

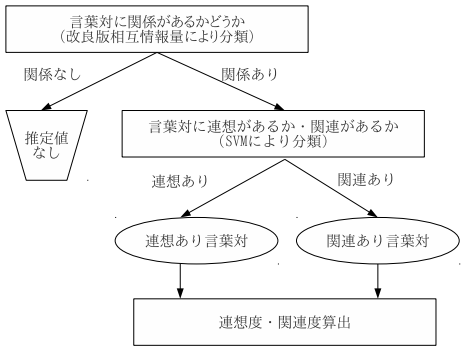

図 1: 提案手法の概要図

3. 提案手法

本研究では，関連度・連想度の指標値算出のため，はじめに予備実験を行った．本章では，予備実験，および，本提案手法について述べる．

3.1 前準備

前準備として，連想度・関連度算出に用いる 2 単語 (計 475 語対) と，指標値算出の評価対象となる 2 単語 (計 30 語対) を含む Web ページ集合と，2 単語の間に出現する文字列を取得した．2 単語を含む Web ページ集合は，Yahoo!検索 API[*1]を用いて取得した．

連絡先: 高尾美代子，豊橋技術科学大学，takao@la.cs.tut.ac.jp

*1　Yahoo!デベロッパーネットワーク，
http://developer.yahoo.co.jp/webapi/search/websearch

3.2 予備実験

予備実験における，指標値算出手順を以下に示す.

【予備実験における算出手順】
Step 1. 入力として，指標値算出の対象となる単語を，関連があるかどうかを分類する関連分類器，連想できるかどうかを分類する連想分類器の両方に与える.
Step 2. 関連分類器において，関連ありと判断された 2 単語に対し，関連度を算出する.
Step 3. 連想分類器において，連想ありと判断された 2 単語に対し，連想度を算出する.

各分類器は SVM を用いて作成する．SVM 実装としては tinySVM[*2]を使用した．また，2 単語の連想度･関連度は，SVM で求まった分離超平面から，2 単語のテストデータが位置する座標までの距離とした．SVM の素性には単語間に出現する文字列，素性値には (1) で求める確率値を使用した.

$$p(a,p,b) = \frac{f(a,p,b)}{\sum_{p \in Sp} f(a,p,b)} \quad (1)$$

ここで,
$f(a,p,b)$: 単語 a，b をともに含む文書集合において，語 a，b の間に素性 p が出現している文数.
Sp : すべての素性集合.

予備実験における算出手順に従い，単語の分類実験を行ったところ，分類精度は関連分類器で 60.0 %，連想分類器で 30.0 %となった.

以上のように，連想分類器での分類精度が著しく低い結果となった．そこで，分類精度が低い原因の調査を行った．調査としては，連想分類器における，学習済み素性の重みを算出した．連想分類器における，学習済み素性の重みの一部を表 1 に示す.

表 1: 学習済み素性の重み（一部）

連想あり語対の素性	重み	連想なし語対の素性	重み
の	0.2774	ではなく	-0.00014
（	0.08983	スイッチ、	-0.00014
・	0.06975	」、「	-0.00025
と	0.01938	対応	-0.00028

以上のように，連想分類器においては,「連想なし」語対に対する，素性の重みがほぼ一定値となっている．このことから,「連想なし」語対において，特徴がないと言える.

SVM は，それぞれが特徴を持った 2 つのクラスへの分類器なので，一方が特徴のないクラスの場合は分類に適さないと言われている．そのため，連想分類器においては，分類精度が著しく低い結果となったと考えられる.

そこで，分類精度向上のため，分類クラスの双方に特徴を持ったクラスを使用する．ここでは，2 単語を，連想ありクラス，関連ありクラスに分類する.

しかしながら，2 単語間に連想関係，関連関係の両方が存在しないという，2 単語間に関係がない場合が考えられる．関係がない 2 単語に指標値を付与することは，メニュー階層評価において大きな問題となる．そのため，入力として与えられた 2 単語から，はじめに関係なし語対を除去し，その後，連想ありクラス，関連ありクラスに分類を行う手法を提案する.

3.3 提案手法

提案手法の指標値算出手順を以下に示す.

【提案手法における算出手順】
Step 1. 指標値を付与したい 2 単語に関係があるかどうか分類する（以下，関係分類ステップ）.
Step 2. Step 1 で関係があると判断された 2 単語に対し，連想があるか，それとも関連があるかに分類する連想・関連分類器において分類を行う.
Step 3. 連想・関連分類器において，連想ありと判断された 2 単語に対し，連想度を算出し，関連ありと判断された 2 単語に対し関連度を算出する.

Step 1 の関係分類ステップについて詳細に述べる．関係分類ステップでは，改良された Pointwise Mutual Information（以下，補正 PMI）[3] を使用し，次の手順により関係があるかどうかを分類する.

【関係分類ステップにおける分類手順】
Step 1. 単語 a，b に対し，補正 $PMI(a,b)$ を求める.
Step 2. 補正 $PMI(a,b)$ がしきい値 t[*3]以上であれば関係あり，t 未満であれば関係なしと判断する.

なお，補正 $PMI(a,b) = \frac{P(a,b)}{P(a)P(b)} \frac{f(a,b)}{f(a,b)+1} \frac{min(f(a),f(b))}{min(f(a),f(b))+1}$ で求める.
ここで，$P(a,b) = \frac{f(a,b)}{N(a,b)}$ であり,
$f(ab)$：語 a，b を共に含む文書集合において，語 a，b を共に含む文数
$N(a,b)$：語 a，b を共に含むデータ集合における，すべての文数
また，$P(a) = \frac{f(a)}{N(a,b)}$，$P(b) = \frac{f(b)}{N(a,b)}$ により求める.

Step 2 で用いる各分類器は SVM を用いて作成し，SVM としては tinySVM を使用した．また，連想度・関連度は SVM で求まった分離超平面からの，2 単語に対する座標までの距離とした．SVM に用いた素性，素性値については予備実験時と同様である.

4. 評価実験

提案手法における算出手順に従い，単語の分類実験を行った．分類実験に使用した学習データを以下に示す．テストデータは，心理値が付与されている語対 30 語を含む Web ページ集合を用いた.

学習データ
【連想・関連分類器】
正例：関連あり語対 123 語を含む Web ページ集合
負例：連想あり語対 77 語を含む Web ページ集合

*2 TinySVM, http://www.chasen.org/~taku/software/TinySVM/

*3 しきい値 t は心理値を参考にして人手により決定する.

4.1 実験結果

分類精度は，関係分類ステップで 80 %，連想・関連分類器では 90.47 %となった．また，算出した指標値を評価するため，指標値と心理値とのピアソンの積率相関係数を求めた．相関は，図 2，図 3 のようになった．

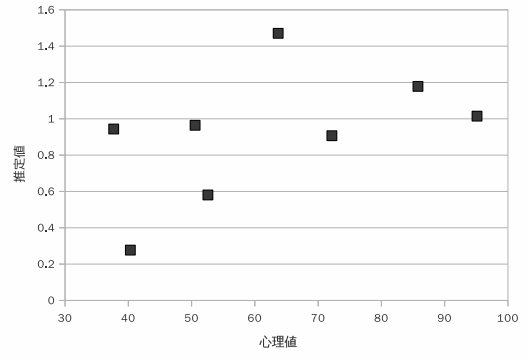

図 2: 関連度における心理値と指標値の相関 (相関係数:0.49)

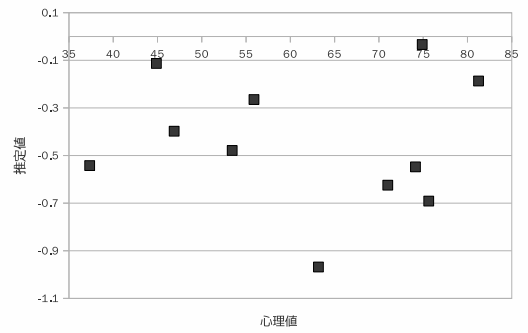

図 3: 連想度における心理値と指標値の相関 (相関係数:-0.03)

5. 提案手法 2

予備実験に比べ，提案手法では分類精度を向上させることができた．しかしながら，心理値と算出した推定値との相関は十分でないと考えられる．

そこで，言葉対を連想・関連関係とすることの適切さを推定することを提案する． これは，例えば，提案手法の Step 2 において,「風量-マイナスイオン」が連想関係であり，かつ,「風量-吹き出し口」が連想関係であると判定された場合において，メニュー階層において,「風量」の子として「マイナスイオン」と「吹き出し口」のどちらを選択する方が適切かを決定するためには，連想・関連関係 にするべき適切さの大小を算出すれば良いという考えに基づくものである．

ここでは，連想関係・関連関係の判定には提案手法を用い，連想・関連関係とするべき適切さを，補正 PMI により算出する．

6. 評価実験 2

連想・関連関係とするべき適切さを，提案手法 1 において，連想あり，もしくは，関連ありと判定された 2 単語に対して，補正 PMI により算出した．また，算出した指標値を評価するため，指標値と心理値とのスピアマンの順位相関係数を求めた．

6.1 実験結果

指標値と心理値とのスピアマンの順位相関は，図 4, 図 5 のようになった．

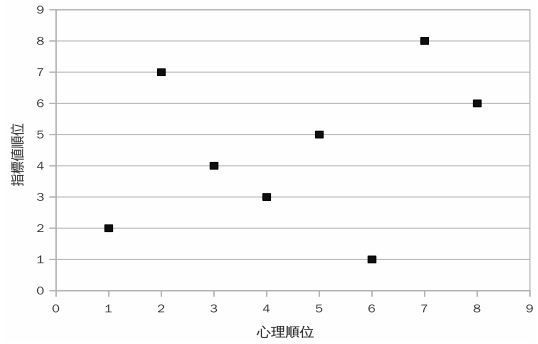

図 4: 関連度における心理値と指標値の順位相関
(順位相関係数：0.31)

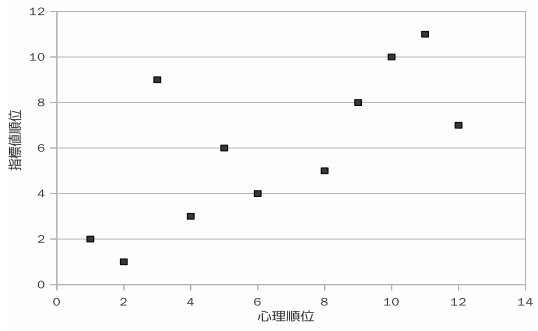

図 5: 連想度における心理値と指標値の順位相関
(順位相関係数：0.70)

連想・関連関係にするべき適切さの大小を算出することにより，連想度においては順位相関係数が 0.7 というやや高い相関を得ることができた．しかしながら，連想度・関連度共に，心理値と算出した推定値との相関はまだ十分とは言えない．

そこで，関連度において，心理順位と指標値順位が大きく異なっている語対に対し，語対が共起している文（以下，共起文）の特性を調査した．今回，心理順位と指標値順位が大きく異なっていた語対は「コントラスト-明るさ」と「電話番号-住所」の 2 つである.「コントラスト-明るさ」は，心理順位が低いにも関わらず，指標値順位が高く,「電話番号-住所」は，心理順位が高いにも関わらず，指標値順位が低くなっていた語対である．

「コントラスト-明るさ」,「電話番号-住所」の共起文例を以

下に示す.

「コントラスト-明るさ」の共起文例
コントラスト 明るさを変える
コントラスト、明るさ

「電話番号-住所」の共起文例
携帯電話番号 住所検索 住所調査 住所調べる
電話番号住所検索地図
電話番号から住所検索

「コントラスト-明るさ」「電話番号-住所」の両方において,「単語 (空白文字列) 単語」というパターンが存在していた．これは，Web ページから 1 文を抽出する際に，タグをすべて除去しているためだと考える．例えば，図 6 に示すような Web ページの場合，タグをすべて除去すると,「ナビ　メモリ地点」となる．

図 6: Web ページ例

つまり，Web ページ上では 1 文で共起していないにも関わらず，共起していると数え上げてしまう．そこで，タグはすべて除去するのではなく，タグの改行変換を行うべきだと考える．これにより,「電話番号　住所検索」というパターンだけではなく,「電話番号住所検索地図」など単語列が羅列しているパターンに対しても，適切な文字列位置で改行が行われ，正しく共起回数を数え上げられると考える．

また,「コントラスト-明るさ」においては,「コントラスト、明るさ」という同じ文が何度も出現していた．これは，広告などで,「コントラスト-明るさ」が含まれる Web ページ集合の中で，類似した Web ページが多く含まれていたためだと考える．そこで，Web ページの類似度を測定し，類似ページは同一 Web ページと見積もる，もしくは，同一文が何度も出現した場合には共起回数を減らすなど，何らかの工夫が必要だと考える．

7. むすび

本論文では，メニュー階層わかりやすさ評価のための，語対の連想関係，関連関係自動判定について検討した．結果，単語間の相互情報量を用いて，あらかじめ，語対に関係があるかどうか分類を行い，関係ありと判断された語対に対してのみ，SVM による機械学習によって連想・関連関係の分類判定を行うことで，分類精度 80 %で判定を行うことができた．

また，補正 PMI を推定値とした場合，関連度における心理値と推定値の順位相関は 0.31，連想度における心理値と推定値の順位相関は 0.70 となった．

以上より，語対の連想関係，関連関係の自動判定は，単語間の補正 PMI 値と SVM による機械学習を用いることで，高精度に判定することができた．また，連想関係，関連関係の語対の適切さの推定は，補正 PMI を推定値とし，推定値の大小を用いて推定できる可能性を示すことができた．

今後は，大量の学習データおよびテストデータを用いて，2 単語の連想度・関連度の指標値を算出すると共に，心理値と指標値の相関が低い原因を考察し，改善を行う予定である．

参考文献

[1] 岡本潤，石崎俊，”概念間距離の定式化と既存電子化辞書との比較 ”，自然言語処理，vol.8，No.4，pp.37-54，2001.

[2] 北島宗雄，高木英明，山本哲生，張勇兵，”潜在意味解析（LSA）を利用した Markov 連鎖モデルによる階層メニュー探索過程の評価”，情報処理学会論文誌，vol.43，No.12，pp.3722-3732，2002.

[3] Patrick Pantel and Deepak Ravichandran, Automatically Labeling Semantic Classes, Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics, 2004.