技術書の索引から抽出した用語関係データに基づく技術用語マッピング手法の提案

猪口真吾
Shingo INOGUCHI
情報メディア学専攻修士課程　1年

1. はじめに

現在，Wikipedia(注1)や Qiita(注2)，SlideShare(注3)など，知識を共有するためのさまざまなサービスがある．こうしたサービスには基本的にはキーワード検索機能が提供されている．ある技術について調べたいとき，技術用語でキーワード検索することで，関連する情報を得ることができる．しかしながら，ある技術分野における初学者にとっては，どの技術用語で検索をかけるべきか，その選定が難しい．適切な情報を得るためには，どの技術用語とどの技術用語とか関連するのか把握しておく必要がある．

このような問題に対し，本研究では，技術用語間の関係性を可視化するインタフェースを検討する．技術用語間の関係性を抽出する基となる情報源としては技術書の索引に着目する．技術書の索引には，その技術書における索引語（技術用語）の出現ページが一覧としてまとめられている．この索引から技術書－索引語（技術用語）－出現ページの関係データを抽出する．ここで，同一技術書において出現ページが近い技術用語は互いに関連性が強いと仮定することで，技術用語間の類似性を定義する．この類似性を基に，技術用語の関係性を 2 次元空間に可視化したインタフェースを構築する．

2. システム概要

本章では，提案手法である技術用語マッピング手法について説明する．

図 1 は，提案システムのインタフェース，図 2 は，システム構成である．インタフェースは，検索ビューとマップビューから構成される．

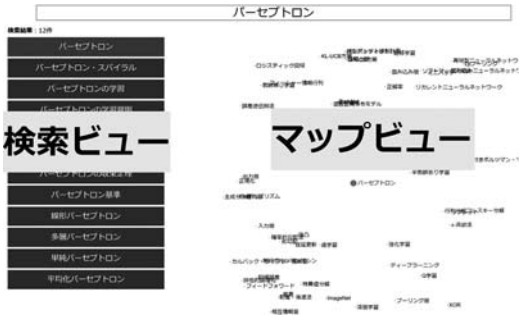

図 1　システムインタフェース

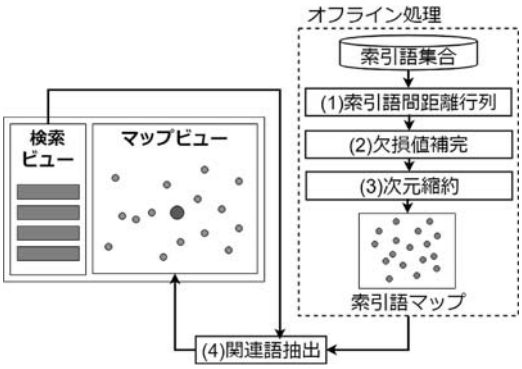

図 2　システム構成

ユーザは検索ビューから任意のキーワードで索引語を検索することができる．マップビューには，検索された索引語周辺の索引語マップが表示される．ユーザは，このマップビューから興味のある索引語に関連する索引語を探索し，調べることができる．

このようなシステムを実現するためには，あらかじめ索引語マップを用意しておく必要がある．本研究では，技術書から索引語集合を抽出し，技術書内における索引語間のページ差に基づき，索引語間距離行列を作成する．この索引語間距離行列を基に，多次元尺度構成法により，2 次元空間による索引語マップを作成する．

3. 評価

3.1　データセット

本実験では，35 冊の技術書を対象とし，各技術書の索引から人手により索引語を抽出し，データベ

ースに登録した．ユニークな索引語数は 5,263 であった．また，登録されたタプル数は 9,072 であった．

3.2 定性分析

索引後マップの作成において，索引語間距離行列上に欠損値の多い索引語同士が塊となってマッピングされる問題がある．索引語マップにおいて同一縮尺で塊となっている索引語とそうでない索引語を表示すると，塊となっている索引語が視認しづらくなる．そこで，DBSCAN を用いて索引語マップをクラスタリングすることにより，索引語同士の塊を判別する．

クラスタリングの結果，全索引語（5,263 語）の内，約 84%（4,416 語）がいずれかのクラスタに分類された．

検索語に対して提示される関連語についての検証を以下の 2 パターンで行う．

（1）どのクラスタにも属さない独立した索引語（DBSCAN における外れ値）の関連語．但し，いずれかのクラスタに属す索引語は関連語から排除．

（2）クラスタに分類された索引語と同一クラスタの関連語（距離行列において欠損値が多い索引語）

1 の場合，検索語：パーセプトロンに対し，

誤差逆伝搬法

入力層

出力層

などの関連語が提示され，また，関連語同士の距離も正しく表示している．例示すると，パーセプトロンと入力層・出力層はそれぞれ関連するが，パーセプトロンと入力層・出力層との関連度と比較すると，入力層と出力層の関連度の方が高い．また，提示された関連語に関して，極端に関係しない用語は確認できない．これは，用いたデータセットが特定の分野に集中していたためである可能性がある．

2 の場合，検索語：音声区間検出に対し，

SNR（信号対雑音比）

音声パワー

ゼロ交差率

などの関連語を提示できている．しかし，1 冊の技術書にしか掲載されていない索引語は，索引語間距離行列において欠損値が多いため，他の技術書の索引語との位置関係の算出が困難である．また，そのような索引語の場合，同一クラスタ内での位置関係は単純に索引語のページ順となるため，マップとして表示する利点は乏しいと考えられる．

4. おわりに

本研究では，技術用語の関係性を可視化するインタフェースの実現に向けて，技術用語マッピング手法を提案した．提案手法により，索引語間距離を算出可能な技術用語間において提示される関連語の妥当性を確認した．また，提案手法では索引語間距離を算出不可能な技術用語が索引語マップ全体に対して大きな影響を与えていることを確認した．

今後，提案手法で算出不可能な索引語間距離の扱いについて検討する．

注

（注 1）：https : //ja.wikipedia.org/

（注 2）：https : //qiita.com/

（注 3）：https : //www.slideshare.net/