2N-05

タグデータの校正のためのデータ間の共起関係に基づく関連タグの抽出に関する研究

石田　祐也†、廣重　法道†、鶴田　直之†
福岡大学工学部電子情報工学科†

1. はじめに

大規模データベース（DB）を構築するにあたって、人手に頼ったデータ入力に対する信頼性と検索効率の担保が課題になる。この課題の原因のうち、例えば専門用語の同義語は、辞書化して事前に入力規則を決めることが難しく、入力時に対策することが難しい。一方、一旦大量のデータが登録されてしまえば、付随情報の類似性から、同じ意味を表している言葉の候補を抽出することができ、校正や対策の素案を作ることができる。そこで、本稿では、データ間の共起関係に基づき関連データを抽出する手法について検討し、道路維持管理業務データに適用した。また、地質コンサルタントに抽出結果を評価してもらい、その有効性を検討した。

2. 大規模 DB のデータ入力の課題

2.1. データ入力時の統一化

DB は、データ入力時に人手を介する場合、誤入力による信頼性の低下と、表現の多様性による検索効率の低下の可能性がある。そこで、データ入力の正確性と統一性を高めるために、何らかの入力規則を定め、規則通りにしか入力できなくするか、規則違反の入力を検出して再入力を求める工夫がなされるのが通例である。具体的には、次のような方法が考えられる。

- 入力漏れ：項目を必須項目に指定して防止
- タイプミスや表記のゆらぎ：入力内容をメニューから選ぶようにして防止
- 同義語：同義語を辞書化して画一的な表現に自動変換

しかし、DB が大規模化すると、値（キーワードや単語）の種類が増える。また、複数の異なる領域に跨るデータを収集する場合は、同義語の辞書化にも困難が生じる。これらの要因によって上記の対策が困難になると予想される。

Study on Extraction of Related Tags Based on Co-occurrence Between Data for Tag Correction
†Y. Ishida, N. Hiroshige, N. Tsuruta・　Fukuoka University

2.2. 道路維持管理業務データの例

筆者らは、道路維持管理業務データと地質学のデータを連携させた大規模 DB を構築し、防災・減災に役立てる研究[1]を進めている。九州地区のある道路維持管理業務データ集合を目視で調べてわかった不揃いの例を示す。

- 入力漏れや、かな漢字変換ミス
- 「九州北部豪雨」に対する多様な表現：7 月における豪雨、降り続いた集中豪雨、九州北部を襲った豪雨など
- 専門英語のカタカナ表記と日本語表現の混在

3. 提案手法

3.1. 概要

2 章で述べた課題をデータ入力時にすべて解消するのは困難である。一方、DB が大規模であれば、既に入力済みのデータを分析することにより、入力ミスの傾向や表現のゆらぎと言った、そのデータベース特有の不揃いを検出し、入力規則を改善できる可能性がある。

そこで、本稿では、登録データに対し、単語間の共起関係を分析し、共起率の高い単語を、字面では判断がつかない入力ミスや表現のゆらぎの候補として抽出し、専門家により不揃いの修正や辞書化を促す方法を提案する。

3.2. 処理の流れ

提案手法の処理の流れを図 1 に示し、主要な部分について述べる。形態素解析は、文章を単語に分割するために用いる。ここで、複合語になっている専門用語（例えば「吹付法枠工」や「自然石積護岸」）が分割されないように土木用語を辞書に登録する必要があるが、本稿の執筆時点では未実装である。

共起行列は、{単語}×{管理業務}の表になっており、管理業務ごとのデータに各単語が何回ずつ出現したかを値として持たせたものである。

潜在的意味解析[2]では、特異値分解を用いて共起行列を圧縮する。これにより、雑音の除去

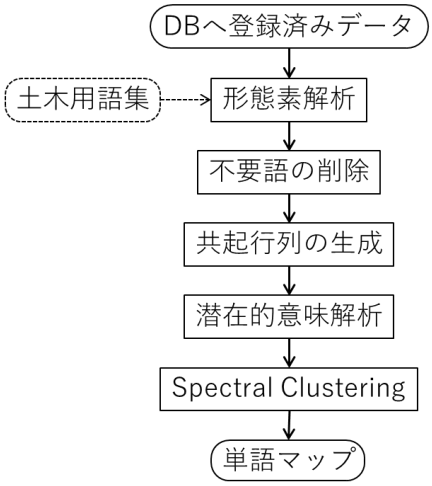
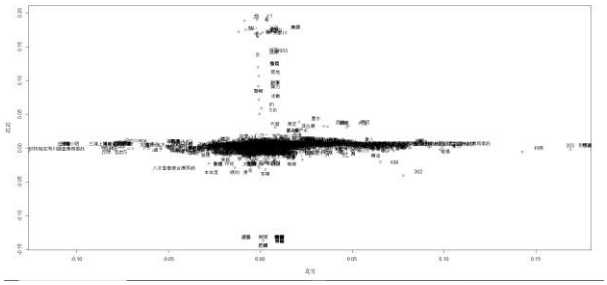


図 1　処理の流れ（破線部は未実装）

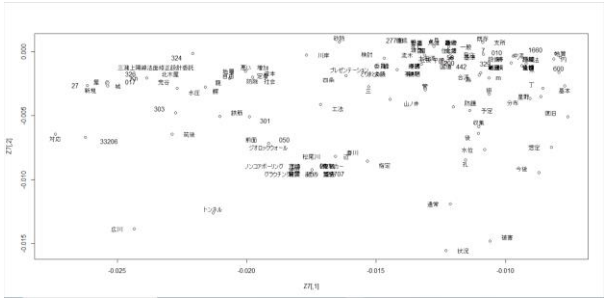
効果に加え、共起性の強い単語同士（あるいは類似した文脈の管理業務）が似たベクトルで表現されるようになることを期待している。

Spectral Clustering[3]では、単語のベクトルの似た者同士が連続するように単語の並べ替えを行う。ここで得られる２種類の並べ替え結果を横軸と縦軸にして単語の関連度を距離で表したものが単語マップである。

以上までは自動化する。最終的な表現のゆらぎと同義語の発見は、単語マップにより関連性の強い単語群を専門家により分析して検出する。そのために、注目単語の周辺だけを拡大するツールと単一の項目に含まれる単語のみを表示するツールを作成した。図 2 に全体の単語マップと拡大マップの例を示す。



(a)全体の単語マップ



(b)拡大マップ

図 2　単語マップの例

4. 実験

4.1. 概要

2.2 節で述べた道路維持管理業務データ集合に対して提案手法を適用し、専門英語のカタカナ表記に共起性の高い日本語を関連語の候補として道路地盤コンサルタントに確認してもらった。

管理業務データは、点検等の業務と施工を伴う工事に大別され、合わせて 212 件からなる。各管理業務は、(名称、概要、(業務概要 ｜ 工事内容))の３項目からなる。

4.2. 結果

8 つのカタカナ表記のうち、専門家によって関連性が認められた 5 つの共起性の強かった語を表 1 にまとめた。有効な関連語が得られなかった 3 つは、カットオフ、ジオテキスタイル、ジオロックウォールであった。得られた高共起性語はカタカナ表記と共起性が高いと同時に他には見られない珍しい語であった。この意味では、DB の大きさが不十分であるとも言える。

表 1　関連が抽出できた共起性の強い語

カタカナ表記	高共起性語	関連性
ボックスカルバート	水桝	関連大
グラウチング	充填	やや同義
パッカー	充填	関連大
アスカーブ	側道	関連大
プレテン	ひび	関連大

5. おわりに

登録済みデータ間の共起関係に基づき関連データを抽出し、データの不揃いを改善する手法について検討し、道路維持管理業務データに適用した。英単語と日本語の同義性に基づく関連語を抽出でき検索効率の向上が見込まれた。今後は、より大きな DB の構築と検証が必要である。

[参考文献]
[1]矢部,他、道路法面点検データの公開に向けたブロックチェーンを用いたデータの信憑性担保の研究、情報処理学会第 80 回全国大会
[2] S. Deerwester, etc, Indexing by Latent Semantic Analysis, Journal of the American Society for Information Science, 41(6):391-407, 1990.
[3] S. Guattery, etc, On the performance of spectral graph partitioning methods, Annual ACM-SIAM Symposium on Discrete Algorithms, 1995