WordNet および Wikipedia と連携するソースコード上の関連語マイニングツール

山下 大貴[1]　竹内 和広[1]

概要：ソースコード上で使用される自然言語は，プログラム中で慣用的に使われる語，アプリケーション固有の語など，一般的な英単語と専門的な語との組み合わせで使われることが多い．プログラム上の語は一般的な語の意味とは異なる専門的な語用が存在しており，ソースコード上の専門的な語とその関係を扱う言語資源の構築は現在研究課題となっている．このような背景から，我々はソースコード分析用の言語資源を構築するために，英語辞書である WordNet と百科事典である Wikipedia の関連付けを意識してソースコード集合に含まれる語間の関連性をマイニングするツールを開発している．

1. はじめに

ソースコード上で使用される自然言語は，プログラム中で慣用的に使われる語，アプリケーション固有の語など，一般的な英単語と専門的な語との組み合わせで使われることが多い．また，関連研究 [1][2] が指摘しているように，ソースコード上の語は一般的な語の意味とは異なる専門的な語用が存在している．他方，ソースコード上の専門的な語とその関係を扱う言語資源の構築は現在研究課題となっている [2]．このような背景から，我々はソースコード分析用の言語資源を構築するために，英語辞書である WordNet[3] と百科事典である Wikipedia[4] の関連付けを意識してソースコード集合に含まれる語の間の関連性をマイニングするツールを開発している．

2. 類義語マイニングツール

2.1 ソースコードに対するマイニング前処理機能

ソースコード上の自然言語を扱うためには，語の抽出処理の共用化が課題となる．我々のツールでは，ソースコード内の以下に示す要素をマイニングの基礎データとして抽出できる機能を持つ．

- クラス名
- メソッド名
- 変数名

ソースコードでは，以上の要素には，キャメルケース・スネークケースといった形式の語の組み合わせで表現される．ツールは，それぞれのケースに対応して処理を行うことができる．

2.2 Randomized SVD

単語の使用は使い手に依存することが多く，それを吸収するためには，”car”と”automobile”など語間の関係（この場合は類義語）を捉える必要がある．自然言語処理で一般的に関連語抽出に用いられる手法の一つに LSI(Latent Semantic Indexing) [5] がある．ソースコード分析でもこの手法の応用は有益と考えられるが，高速かつ高次元の処理が必要となるため，我々は図 1 のように，行と列方向でのサンプリングにより，効率的に次元圧縮を行うことができる手法である Randomized SVD[6][7] を採用したツールを提供する．我々のツールでは，図 1 のように 2.1 節のマイニング前処理と高速・高次元の LSI 処理の処理結果は，任意の数の関連語集合を処理結果として出力する．具体的には，LSI は行列の SVD により得られた対角行列 S に対し，指定した次元に圧縮するが，我々のツールではその処理を内部的に行うため，ユーザはソースコードから，高次元の行列表現を意識することなく，関連語集合を直接得ることができる．

3. WordNet との連携

語間の意味関係の発見は，既存の辞書知識との相関が重要な課題となる．我々のツールでは，ツールで発見した関連語集合中の語間の意味関係を，WordNet と比較できる機能をもつ．WordNet は，一般的な語に関する語間の意味関係が記述されており，このことにより，ソースコード内の語の意味関係が，WordNet の語関係とどの程度違うかを客

[1] 大阪電気通信大学 大学院工学研究科
Osaka Electro-Communication University, Graduate School of Engineering

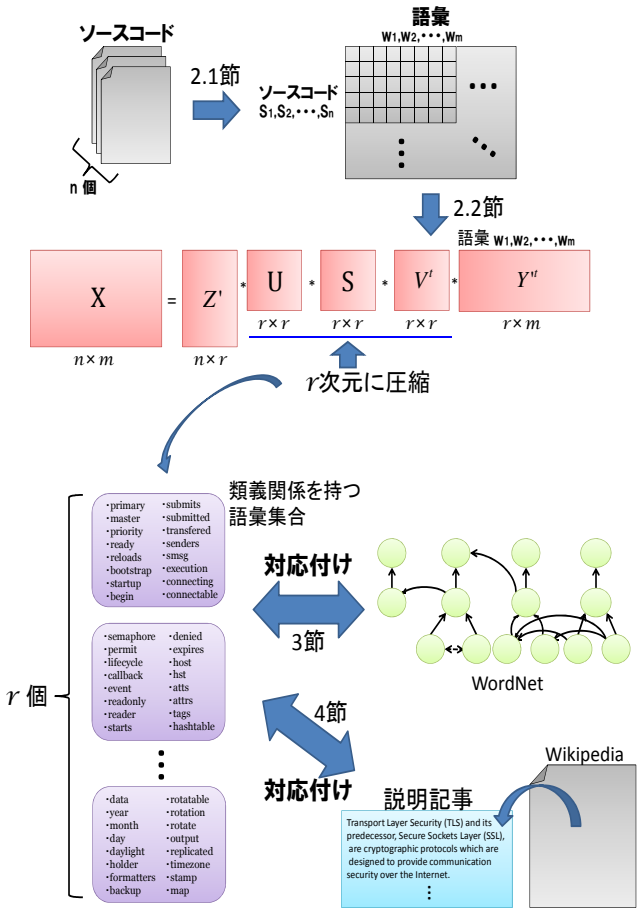

図 1 Randomized SVD 処理手順

観的に比較することができる．

この機能の利点を我々は次のように考えている．LSI では，図 1 の行列 S の大きさを調整することにより，語間の意味的関係を同定する．例えば，1 次元に圧縮した場合，すべての語が入る 1 つの関連語集合が出力される．2 次元に圧縮すると語が 2 つの関連語集合に分けられる．この時，関連語集合の数をいくらに設定するかが課題となるが，以上の機能を利用すれば，WordNet 上で特定の語間関係をもつ語対が同一集合に入った時，というような知識と関係付けた処理が可能となる．

また，未知のソースコード内での語用がどれほど一般的な語用と異なっているかを評価することができる．例えば，プロジェクトにおける語用の規約を WordNet の記法を利用／拡張して定義しておけば，客観的にソースコードの規約適合の度合いを評価することに役立つと考えている．

4. Wikipedia との連携

ソースコード上の語には WordNet には存在しない専門用語も多々用いられる．それらの多くは，そのプログラムの開発対象と強く関連していると思われる．しかし，何を専門用語として定義するかには客観的な整理法が必要であり，我々は，それに Wikipedia を用いることが有益ではないかと考え，開発中のツールでは WordNet だけではなく，ツールによる出力結果と Wikipedia との記事との対応付けを行う機能を作成した．Wikipedia を対応付け先とした理由は，wiki を利用して構築された Web 上の大規模百科事典であり，一般的な概念だけでなく，幅広い対象の記事を含んでおり，プログラムの対象や目的となる記事も含んでいる可能性が高いであろうと考えたからである．

この機能を利用すれば，未知のソースコード群から抽出された関連語集合から，ソースコードの対象領域や処理目的に関する Wikipedia 記事が同定可能であると考えている．我々の予備的実験では，遺伝的アルゴリズムやニューラルネットワークといった専門的プログラムのソースコード群と Wikipedia の当該記事とが自動的に対応付けられることが判明している．Wikipedia 記事は多岐にわたるため，同定先の限定や，対象とするソースコードにも限定が必要である課題はあるが，前節の WordNet との対応付け機能と併せて精緻化していけば，ソースコードとその仕様書などの自動対応付けや，実装の妥当性検証などの応用に有益ではないかと考えている．

5. おわりに

本稿では，我々が現在開発している，WordNet，Wikipedia との連携を意識してソースコード上の類義語関係をマイニングするツールの紹介をした．ツール群はポスター会場でデモを行う予定であり，また，Web 上でも公開していく予定である．ご意見を賜れば幸いである．今後は，このツール群を使って，プログラム実装に関わる文書との対応付けをより精緻にする研究を進め，より高度なリポジトリマイニングに寄与する基盤資源整備に貢献していきたい．

参考文献

[1] M. J. Howard, S. Gupta, L. Pollock, and K. Vijay-Shanker, Automatically Mining Software-Based, Semantically-Similar Words from Comment-Code Mappings, Proceeding MSR '13 Proceedings of the 10th Working Conference on Mining Software Repositories, pp.377-386 (2013)
[2] J. Yang and L. Tan, Inferring Semantically Related Words from Software Context, Mining Software Repositories (MSR), 2012 9th IEEE Working Conference on, pp.161-170 (2012)
[3] WordNet http://wordnet.princeton.edu.
[4] Wikipedia http://www.wikipedia.org/
[5] C. D. Manning, P. Raghavan, and H. Schutze, Introduction to Information Retrieval, Cambridge University Press (2008)
[6] 岡野原 大輔: 全部分文字列のクラスタリングとその応用, 言語処理学会 第 17 回年次大会 発表論文集, pp.65-68 (2011)
[7] N. Halko, P. G. Martinsson, and J. Tropp, Finding Structure with Randomness: Probabilistic Algorithms for Constructing Approximate Matrix Decompositions, Society for Industrial and Applied Mathematics, SIAM REVIEW Vol.53, No.2, pp.217-288 (2011)