3B-6

Ontologyを利用したTraceabilityリンクの評価

田中　勝輝†　　小形　真平†　　海谷　治彦†　　海尻　賢二†

信州大学工学部†

１．はじめに

ソフトウェアドキュメント間のTraceabilityリンクの確立をすることによって、ドキュメント間の関連性を把握することができ、システム開発の保守作業に役立つ。リンクの確立方法として、ドキュメントに出現する単語の共起性に基づくIR手法(情報検索手法)がよく使われている。このIR手法を用いたTraceabilityリンクの確立の精度を向上させるために、数多くの研究で多くの手法が挙げられている[1]が、ドキュメントの特徴の側面から見た有効な手法については考察されていない。よって、開発者が作成するソフトウェアドキュメントにも多種多様な書式・形式があるため、数多くの精度向上手法の中でそのドキュメントにより有効な手法を見つけ出すのは困難である。そこで、精度向上手法の一種である出現単語の同義語・類語に関するontologyを考慮したIR手法を用い、数十種類のドキュメントに対して、ontologyの有効性を示し、ontologyが有効なドキュメントの特徴を探っていく。本研究では、IR手法におけるWordNetを利用したOntologyの取り入れ手法を提案すると共に、Ontology手法が有効なドキュメントを判別する手法も提案する。

２．実験

Ontologyを利用したTraceabilityリンクの有効性をドキュメントの側面から示すために、ドキュメントのOntologyに対する特徴量を設定し、Ontologyの有効性を示す精度と合わせてマイニング実験を行う。

２．１　対象ドキュメント

正しいリンクが予め分かっている参照データセットを使用する。ドキュメントの特徴に焦点を当てるので、10種類のソフトウェアプロダクト群の中から、49種類のドキュメントペアを用意した。使用されている自然言語はすべて英語で、中にはJavaも含まれている。

これらのドキュメントペアには、要求仕様書とソースコード、ユースケースとテストケースなどの組み合わせがある。また、予め小項目ごとファイルが分割されている。これらのソフトウェアプロダクト群のほとんどは、Center of Excellence for Software Traceability[2]で提供されているプロダクト群であり、Traceabilityリンクの研究の実験対象によく使われ、評価されているドキュメント群である。

２．２ IR手法上でのOntologyの取り入れ方

IR手法を用いたTraceabilityリンクの確立には、図１のような各ドキュメントと出現単語の関係を表したterm-document行列が使われる。この行列を元にコサイン類似度で各ドキュメントの類似度が計算される。しかし、図１では”trip”と”travel”が同義語ということが考慮されていない。従って、”trip”と”travel”が同義語ということ反映させるために、図１の括弧内に示すような値を補正値として与える補正term-document行列を作成する。

	d1	d2	d3	d4	d5	d6
ship	1	0	2	0	0	0
boat	0	4	0	0	0	0
ocean	2	1	0	0	0	0
voyage	1	0	0	1	1	0
trip	0 (1)	0 (0.5)	0	3	0	5
travel	2	1	0	0 (1.5)	0	0 (2.5)

図１（補正)term-document行列

図１のd1に”travel”という単語が出現しているが、”trip”という単語が出現していないので補正値としてtravelの出現回数の半分の重みを与えた。本研究では、この補正term-document行列を用いてコサイン類似度計算した値を、ontologyを考慮したドキュメント間の類似度とする。本研究では、ドキュメントペアから出現単語を抽出し、ストップワードを除去し、原形に直した単語群を対象とした。

２．３ WordNet

補正term-document行列の中で、単語の同義語を求める際、本研究ではWordNetを使用した[3]。WordNetとは、英語の概念辞書システムであり、同義語はsynsetと呼ばれる部分に分類されている。これに日本語の概念辞書を追加した日本語WordNetがある。本研究では、この日本語WordNetを使用して、各ドキュメントの出現単語のontologyを構築する。WordNetは、一

Evaluation of Traceability Links with ontology

† Katsuki Tanaka, Shinpei Ogata, Haruhiko Kaiya, Kenji Kaijiri (Faculty of Engineering, Shinshu University)

Copyright ©2013 Information Processing Society of Japan.
All Rights Reserved.

般的な意味合いを持つ辞書であり、特定の分野に特化した辞書ではない。実験対象とするドキュメントは、ソフトウェア分野にそれぞれ特化したものを扱うが、各ドキュメントに対して均一な精度を測るため、ontology 構築に WordNet を用いることにした。この日本語 WordNet データベースの synset 部に、与えた単語の各品詞の意味合いでの同義語リストに問い合わせることができる。例えば、”travel”の同義語をデータベースに問い合せると、 locomotion, traveling, travelling, go, move, locomote, jaunt, trip, journey といった 9 種類の同義語が取得できる。ドキュメントペアの全単語群をこの WordNet のデータベースに通し、同義語群の単語からドキュメントの単語群に含まれている単語をピックアップし、ドキュメントの全単語の同義語リストを作成する。この同義語リストを元に term-document 行列に補正値を与え、同義語を考慮した補正 term-document 行列を生成する。

２．４ Ontology を用いた Traceability リンクの精度評価

IR 手法における Traceability リンクの精度は、Precision と Recall を使って評価される。計算されたドキュメントペアの類似度を閾値と上位 N 個を用いてリンクありと判定する候補リンクとして抽出し、正解リンクと照らし合わせ、図２に沿って Precision と Recall の値を算出する。

候補リンク ＼ 正解リンク	リンクあり	リンクなし
リンクあり	True Positive (TP)	False Negative (FN)
リンクなし	False Positive (FP)	True Negative (TN)

$$\text{Precision} = \frac{TP}{(TP+FP)} \qquad \text{Recall} = \frac{TP}{(TP+FN)}$$

図２ Precision と Recall

本研究では、Ontology を用いた精度値と Ontology を用いないプレーンの精度値を算出する。Ontology を用いることが有効に働くということは、通常では、リンクがあるのに類似度が低く出てしまいリンクなしと判定されてしまう同義語が使われているドキュメントペアをカバーできることとなる。つまり、図２の FN の減少に繋がり、Recall が上昇すると考えられる。従って、Ontology を用いた Traceability リンクが有効だと言える評価基準として Recall を重視する。

２．５ ドキュメントの特徴量

ドキュメントの側面から Ontology を用いた Traceability リンクが有効だと示すため、Ontology に関わるドキュメントの特徴量を設定する。この値と２．４の精度評価値を合わせてデータマイニング実験を行う。ドキュメントの特徴量として、

- ✓ドキュメント A, B 内の同義語の割合
- ✓term-document 行列の、同義語で影響を受ける要素の割合

等を予定している。

これらのデータを独立データ、ontology 利用による精度向上の有無を依存データとして、マイニングツールを通して、ドキュメントの特性から ontology を用いた手法が有効であるかどうかの判断が可能かを調べる。

３．おわりに

今後は、先ず、提案した ontology 手法の有効性を評価する実験を行い、精度の向上が見込めるドキュメントペアがあるかどうかを検証する。また、ontology 手法を適用することで、精度が低下するドキュメントペアもあるのかどうかも検証し、精度の向上があったペアとそうではないペアを比較し、それらを識別できる特徴量が設定できないか調べる。次にマイニング実験を行い、ドキュメントの特性から ontology を用いた手法が有効であるということの予測が可能であるかどうかを考察していく。現時点では、ontology に関する特徴量として何が適切であるかは明確ではないので、実験によって判断する。上記では ontology に関する特徴量のみを挙げたがドキュメント自体の特徴量や、閾値で抽出した候補リンクと上位 N 個で抽出した候補リンクの違いなども考慮する必要がある。さらには、提案した Ontology 取り入れ法が妥当なものであるのかの評価も必要である。実験結果は発表の場で示す予定である。

参考文献

[1] Xiaobo Wang, Guanhui Lai, Chao Liu: Recovering Relationships between Documentation and Source Code based on the Characteristics of Software Engineering, Electronic Notes in Theoretical Computer Science 243 (2009) 121- 137

[2] Center of Excellence for Software Traceability : http://www.coest.org/

[3]上嶋 宏、三浦 孝夫、塩谷 勇：同義語，多義語の考慮による文書分類の精度向上、電子情報通信学会論文誌 D-I Vol.J87-D-I No.2 pp.137-144 2004 年 2 月

Copyright ©2013 Information Processing Society of Japan.
All Rights Reserved.