D-003

セマンティック WEB 技術を用いた技術ドキュメントの類似性検出方法とその評価

柳田　憲士郎　　塚本 享治

東京工科大学大学院バイオ情報メディア研究科

1　はじめに

近年、インターネットの発展に伴い、多くの技術ドキュメントが電子データとして蓄積、公開されるようになった．現在、これらのデータから目的とするドキュメントを取得するにはサーチエンジンによるキーワード検索が広く用いられている．しかし、上記の方法では表記のずれや検索キーワードを連想できないという理由により本来求めている情報に到達できないことがある．

本報告では、セマンティックＷＥＢ技術を用いてインターネット上に公開されている技術用語辞典から用語の関係性に関する情報を抽出し、この抽出された情報を技術ドキュメントの類似性検出に利用した．また、用いる関係性の違いによる類似性検出の精度評価について検討を行った．

2　アプローチ

2.1　既存研究の現状と問題点

類似性検出によって、関連性の高い技術ドキュメントを検索する研究として[1][2]といったものがある．従来の研究では、技術用語をキーワードとして抽出し、その抽出量によって類似性を計るものであり、比較対照ドキュメント間で同一キーワードが出現しなければ、関連性のある分野で高い分野であっても類似性が低いとみなされてしまう問題がある．

2.2　対象技術ドキュメント・技術用語

類似性検出対象ドキュメントとして、大学内で公開されている情報処理技術分野に関するシラバス、パワーポイント授業資料、卒業論文を用いた．技術用語の関係性の抽出は、IT 用語辞典 e-Words[3]から行った．この用語辞典に登録されている用語を類似性検出時に使用する用語対象とした．

2.3　用語関係性データ構築技術及び検索手法

抽出した用語情報はセマンティック WEB 技術として標準化されている OWL[4]形式によってデータ化を行った．用語情報の抽出の際には GRDDL によるコンテンツ変換[5]を用いて HTML で公開されている用語辞典から用語の関係性に関する記述を抽出し、OWL 形式に変換した．

OWL 形式で構築されたデータから類似性検索に必要となるデータの抽出には、RDF 検索クエリ言語である SPARQL[6]を用いた．また、OWL の推論処理実装には Jena[7]、pellet[8]を用いた．

Similarity Extraction among Technical Semantic Documents and its Evaluation
Kenshiro YANAGIDA, Michiharu TSUKAMOTO
Tokyo University of Technology, School of Media Science

3　技術用語関係性データの構築

3.1　関係性情報の抽出対象領域

類似性検出に用いる用語は、e-words ページ内の「索引」及び「分野別索引」で記述されている用語 8967 語（2008 年５月末時点）を対象とした．これら用語は、一つの用語に対して一つの HTML ファイルが情報として提供されている．この HTML ファイル全てに対して GRDDL を用いた変換処理を行った．

3.2　抽出した関係性について

前項で対象とした用語 HTML ファイルに対して、図１の黒枠で囲まれている情報を用語の関係性情報として抽出した．

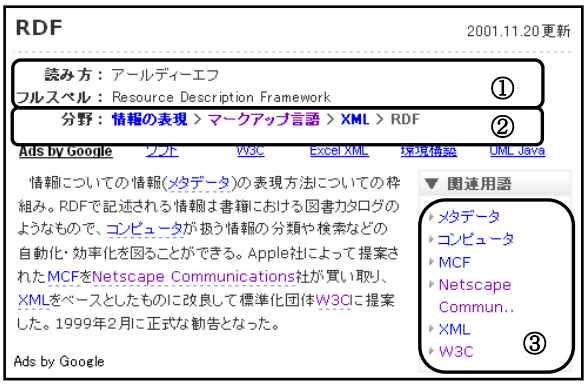

図１　HTML からの関係性情報分類

①は該当用語に関する別の表記方法に関する情報である．この情報は OWL 記述を行う際に「同義語」というプロパティ区分を行い、「読み方」「フルスペル」といった内容を「同義語」として抽出した．②は用語の関係性を階層構造化されたものである．この用語の階層関係を「親用語」「子用語」というプロパティ区分を行い、階層構造の上下関係が分かるように抽出した．③は説明文中にある用語を抽出したものである．ここに列挙されている用語を「関連用語」として抽出した．

抽出された情報は（「RDF（用語）」-「関連用語（プロパティ）」-「メタデータ（用語）」）といった RDF 構造に沿ったデータとしてデータ化される．

3.3　OWL によるプロパティ制約記述

前節で定義したプロパティに対して表１のような制約記述を行った．この制約記述により、pellet による推論処理を行った際に図２のように関係性を補完し、SPARQL によってそれらの関係性を検索可能となった．

表 1　プロパティ制約記述一覧

プロパティ元	OWL 制約表記	プロパティ先
読み方	rdfs:subPropertyOf （継承関係）	同義語
親用語 関連用語	owl:inverseOf （逆関係）	子用語 関連用語参照元
	owl:TransitiveProperty （推移性）	親用語 子用語

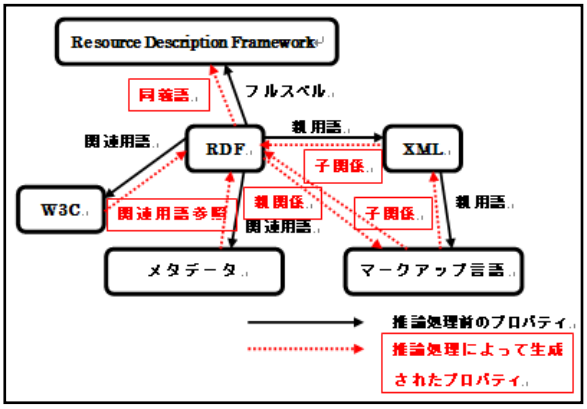

図 2　OWL 制約記述によるプロパティ補完

4　類似性検出

4.1　類似性検出処理手順

以下の処理手順により類似性検出を行う．

(1)　比較対照となる技術ドキュメントに対して、用語関係性データに登録されている用語と一致するものについて全て出現数を抽出する．なお、「セマンティック WEB」といった複数の単語が組み合わさって登録されている用語については、用語の文字数の長い方を優先とし、短い方は抽出しない．（この場合「WEB」という用語は抽出されない）

(2)　抽出した用語を元に SPARQL 文を生成し、pellet による推論処理を行い、抽出用語と何かしらのプロパティで関係が記述されている用語を全て抽出する．

(3)　(1)及び(2)によって抽出した用語を、類似性を判断するためのキーワードとし、これら全てのキーワードに対して抽出した用語間の関係性によって重み付けをし、類似性計算に用いる特徴ベクトルを得る．

(4)　技術ドキュメントごとに得られた特徴ベクトルに対してコサイン値を求めることにより、類似度の値を得る．

4.2　用語間の関係性による重み付け

SPARQL によって得られた用語間の関係性を元に以下の式を利用して類似性計算に用いる文章dのキーワードwに関するスコア$wf(d,w)$を取得する．

$$wf(d,w) = N * \sum_{i=0}^{m} p_{i(w)}$$

ここでの Nはwの出現回数、mは用語間の関係性を表すプロパティ種類の総数を表している．$p_{i(w)}$はキーワードwの関係性p_iで定義された重み付けの値を返す．

4.3　類似性計算手法

前節で得られたスコアを元にベクトル空間法[9]による特徴ベクトルを用いた類似度の計算を行った．文書d_iの特徴ベクトル$\vec{D}_i$は次式のように定義される．

$$\vec{D}_i = [wf(d_i, w_1), wf(d_i, w_2), \cdots wf(d_i, w_n)]$$

この式により得られた二つのドキュメントに関する特徴ベクトル$\vec{D}_i$, $\vec{D}_j$の以下の式を用いてコサイン値を求めることにより類似性の評価が可能となる．

$$\cos(\vec{D_i}, \vec{D_j}) = \frac{D_i \times D_j}{|D_i||D_j|}$$

5　類似性検出評価

大学内で公開されている情報処理技術に関する授業 6 科目のパワーポイント授業資料からテキスト部分のみを抽出し、これら授業資料と以下の関係にある授業の資料との類似性検出を行った．

・担当教員が同じでかつ前提履修が必要又は望ましいとシラバスに記述されている授業資料（類似性高）
・担当教員は違うが「プログラミング」といった同一分野に関する授業資料（類似性中）
・担当教員も分野も別のもの（類似性低）

表 2 は用語間の関係性プロパティ情報の使用可否による類似性適合率の値である．用語間の関係性情報を用いたことにより検出精度が良い結果となった．

表 2　類似性適合率

	関係性情報　○	関係性情報　×
類似性高	83%	66%
類似性中	66%	50%
類似性低	100%	100%

6　まとめ

技術用語辞典で記述されている用語の関係性をセマンティック WEB 技術を用いて利用することにより、類似度検出の精度を上げることができた．今後は複数の技術用語辞典を用いる等の方法により検出精度を上げ、より多くのドキュメントに対して検出評価を行う必要がある．

参考文献

[1]　八太　絵美，文書間の類似度に基づく論文検索システムの開発と評価，日本教育工学会研究報告集 ,JET02(2) pp 91-97,2002.
[2]　丸山崇, 北栄輔, 進化的計算手法を用いたＷｅｂ検索キーワードのクラスタリング手法の提案, 情報処理学会研究報告書 Vol.2006, No.135(20061221) pp. 93-96
[3]　IT 用語辞典 e-Words, http://e-words.jp/
[4]　W3C "OWL Web Ontology Language", http://www.w3.org/TR/owl-features/, 2004
[5]　柳田憲士郎，塚本享治, GRDDL によるコンテンツ変換を用いたオントロジー構築, 2008-DBS-144 pp. 47-54, 2008
[6]　SPARQL Query Language for RDF, W3C, http://www.w3.org/TR/rdf-sparql-query/
[7]　Jena - A Semantic Web Framework for Java, http://jena.sourceforge.net/
[8]　Pellet: The Open Source OWL DL Reasoner, http://pellet.owldl.com/
[9]　Salton, G.,“The Vector Space Model, Automatic Text Processing,“Addison Wesley Publishing, pp.312-325 (1985).