領域オントロジー構築支援環境を用いた オントロジー搭載型検索システムの提案と評価

An Ontology-based Information Retrieval System using a Domain Ontology Development Environment

森田武史[1]　小野穣[1]　洪潤基[1]　川村正則[2]　小出誠二[3]　山口高平[1]

[1]慶應義塾大学
[2]株式会社IHIエスキューブ　[3]株式会社ギャラクシーエクスプレス

要旨

本研究では，オントロジー搭載型検索システム及び領域オントロジーを半自動構築可能なツールである DODDLE-OWL を提案する．DODDLE-OWL により領域オントロジー構築コストを削減する．領域オントロジーを用いた絞込および拡大検索機能を検索システムに導入し，検索精度を向上させる．ケーススタディとして，DODDLE-OWL を用いてロケット運用オントロジーを構築する．また，ロケット運用オントロジーを用いた文書検索実験及びその評価について述べる．

1. はじめに

本研究では，オントロジー搭載型検索システムおよび日本語オントロジーを半自動で構築可能なツールである DODDLE-OWL の提案を行う．より精度の高い検索を実現するために，文書の意味を考慮した意味検索に関する研究が多数行われている．意味検索を実現する方法の一つとしてオントロジーの利用が考えられるが，オントロジーは構築コストが高いという問題がある．本研究では，最初に検索システムに搭載するオントロジー（初期オントロジー）の構築コストを削減するために，DODDLE-OWL を用いる．また，領域オントロジーを検索システムに搭載し，絞込検索および拡大検索により検索精度を向上させる．ケーススタディとして，DODDLE-OWL を用いたロケット運用オントロジーの構築について述べる．また，ロケット運用オントロジーを用いた文書検索実験及びその評価について述べる．

2. オントロジー搭載型検索システム

図 1 にオントロジー搭載型検索システムの全体図を示す．専門文書検索に用いる領域オントロジーを手動で構築するコストは高いため，本研究では初期領域オントロジーを DODDLE-OWL を用いて半自動構築し，検索システムに搭載する．検索システムには，(株)ギャラクシーエクスプレス社(以下 GX 社)で開発された社内文書検索システム GXFinder[2]を用いる．GXFinder はキーワード検索に加えて，領域オントロジーの概念階層を用いた拡大および絞込検索機能，検索した文書がユーザの求める文書かどうかをログとして保存する機能などがある．ユーザが検索作業を繰り返しながら，ログを領域オントロジーに反映させていくことによって，専門文書検索に特化した領域オントロジーを構築できると共に，検索精度の向上も期待できる．本研究の特徴は，検索システムの検索精度の向上と領域オントロジー構築の問題を，領域オントロジー構築ライフサイクルと文書検索を関連付けて解決することである．

3. DODDLE-OWL

DODDLE-OWL[1]は，EDR 電子化辞書[4]と対象領域に関する日本語専門文書を用いて，日本語を概念の表記としてもつ領域オントロジーを半自動で構築可能なツールである．図 2 に DODDLE-OWL のシステムフローを示す．DODDLE-OWL は，入力モジュール，オントロジー構築モジュール，オントロジー洗練モジュール，視覚化モジュール，変換モジュールの 5 つのモジュールから構成される．はじめに，ユーザは入力モジュールにおいて，入力概念を選択する．オントロジー構築モジュールは，オントロジーの基礎となる初期概念階層と概念対のセットを，EDR 電子化辞書と日本語専門文書を参照しながら，入力概念を基に生成する．初期概念階層は IS-A 階層として構築される．概念対のセットは共起性に基づく統計処理を用いて獲得される．これらの概念対の中から特に重要な概念対をユーザが選択し，その間

の関係を定義したものが概念定義となる．オントロジー構築モジュールで構築された初期オントロジーをオントロジー洗練モジュールはユーザとやりとりしながら洗練していく．初期オントロジーを洗練するために，概念変動と概念対のセットの評価の管理を支援する．初期概念階層は一般的なオントロジーから生成されるため，ユーザは概念変動と呼ばれる問題を考慮しながら，初期概念階層を特定の領域に調整する必要がある．それは，特定の部分の概念が領域によって変化することを意味する．概念変動管理のために，DODDLE-OWL は照合結果分析と剪定結果分析の 2 つの戦略を適用する．オントロジー構築モジュールで生成された概念対のセットから重要概念対を評価するための指標として，WordSpace 法による文脈類似度と相関ルールによる信頼度の 2 つの共起性に基づく統計的な手法を用いた重要概念対の評価方法を用いている．構築されたオントロジーは変換モジュールによって OWL 形式に変換される．

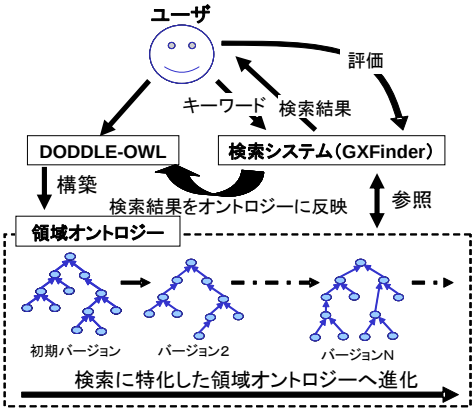

図 1　オントロジー搭載型検索システム

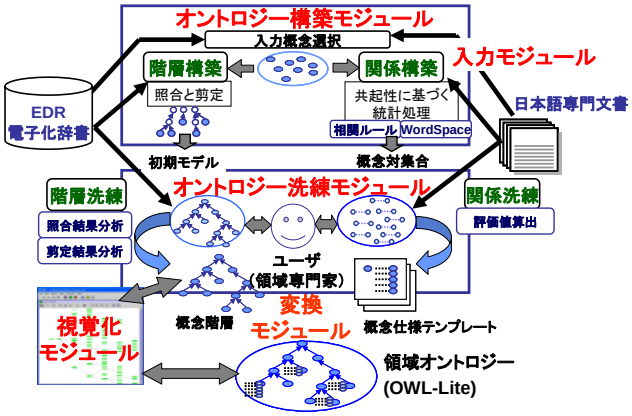

図 2　DODDLE-OWL のシステムフロー

4.　領域オントロジーを用いた検索

GXFinder は，DODDLE-OWL により構築された OWL 形式の領域オントロジーを搭載し，概念階層を利用した絞込および拡大検索を行うことができる．絞込検索は，検索結果数が膨大な場合に検索結果数を絞込み，ユーザの目的に合った文書を見つけやすくする．検索キーワードを表記として持つ概念の下位概念について，それぞれの概念表記を OR で結合し検索を行う．何階層下まで展開するかはユーザが指定することができる．拡大検索は，検索結果数が少ない場合や検索したいキーワードをユーザが漠然としか思い浮かばない場合に，関連する文書を多くユーザに提示し，目的に合った文書を見つけやすくする．拡大検索は，検索キーワードを表記として持つ概念の上位概念および兄弟概念について，それぞれの概念表記を OR で結合し検索を行う．絞込検索と同様に，何階層上まで展開するかはユーザが指定することが可能である．

5.　ケーススタディ

本ケーススタディの目的は，領域オントロジーを用いた拡大検索および絞込検索の有用性を確かめることである．本ケーススタディは GX 社の協力の下，ロケット運用に関する領域オントロジーを構築し，それを GXFinder に搭載し，検索実験を行った．以下では，DODDLE-OWL を用いたロケット運用オントロジーの構築およびキーワード検索と領域オントロジーを用いた検索の比較実験について述べる．また，実験の考察について述べる．

5.1. ロケット運用オントロジーの構築

ロケット運用オントロジーの構築は，1 人のユーザが DODDLE-OWL を用いて約 30 時間かけて，1. キーワード抽出，2. 不要語の削除，3. 多義性解消，4. 階層構築の手順で行った．

1．キーワード抽出． GX 社豊洲分室で作成されたロケット運用に関する 2845 の日本語文書から形態素解析システム Sen[3]によって一般名詞 32,415 語およびその他の名詞 4,470 語を抽出した．名詞が連続

している場合，それらをまとめて名詞の複合語とした．

２．不要語の削除． Sen により抽出された語の中には不要な語が含まれている．DODDLE-OWL に入力する前に 1 字の語，20 字以上の語，文字化けしたと思われる語については，プログラムにより自動的に削除した．切られ方の誤った語および領域にとって不要な語については，ユーザが手動で削除した．最終的に 32,814 語を DODDLE-OWL の入力とした．

３．多義性解消．単語は複数の意味を持つ場合があるため，ある単語を表記としてもつ概念が複数存在する．DODDLE-OWL の入力モジュールでは，入力単語とそれに対応する EDR 電子化辞書中の概念の候補を提示する．ユーザは入力単語に対応する，領域にとって最も適切な概念を選択する．大部分の複合語は，それを表記として持つ概念が EDR 電子化辞書中に存在しない．DODDLE-OWL では，部分照合を行うことによって，多くの複合語の多義性解消を可能にしている．部分照合は，入力単語と EDR 電子化辞書中の概念が持つ表記が部分的に一致することを意味する．完全照合しなかった入力単語については，Sen を用いて形態素解析を行い，先頭の単語を順に除いて EDR 電子化辞書中の概念と対応付けを試みて，最長一致した単語に対応する概念と対応付けを行う．その際に，部分照合概念を照合した概念の下位概念とするか，同義語とするかをユーザは選択可能である．本ケーススタディでは，概念数が膨大であったことから，部分照合した単語については，すべて照合した概念の下位概念とした．本ケーススタディでは，完全照合単語数 4,982 語，部分照合単語数 26,835 語，未照合単語数 997 語となった．

４．階層構築． 階層構築は DODDLE-OWL が自動で行う．DODDLE-OWL は，多義性解消時に部分照合した複合語について，語尾および語頭による階層化を行う．語尾が等しい複合語は，兄弟概念として階層化される．部分照合した入力単語の語尾以前（照合しなかった部分）の文字列を表記として持つ概念が構築中の領域オントロジー内に存在する場合，その上位概念と入力単語の語尾を組み合わせた概念を入力単語の上位概念として定義する．例えば，「計器」の下位概念に「レーダ」，「センサー」，「ゲージ」という概念が定義されているとする．「モデル情報」，「レーダ情報」，「センサー情報」，「ゲージ情報」という複合語を階層化する場合，語尾による階層化では，「情報」の下位概念に，「モデル情報」，「レーダ情報」，「センサー情報」，「ゲージ情報」が定義される．ここで，複合語の語尾以前の単語である，「レーダ」「センサー」「ゲージ」については，領域オントロジー中に共通の上位概念として「計器」が定義されている．よって，「計器」と複合語の語尾の「情報」を組み合わせて，「計器情報」概念を作成し，その下位概念に「レーダ情報」，「センサー情報」，「ゲージ情報」を再定義する．これにより，「モデル情報」と「レーダ情報」，「センサー情報」，「ゲージ情報」という計器に関する情報を分類することができる．本ケーススタディでは，語尾および語頭による階層化により 34,451 概念の初期概念階層が構築された．

本来は，階層構築後，領域に特化した形に階層を修正すべきである．しかし，本ケーススタディでは，ユーザが初期領域オントロジーを洗練するために十分な領域に関する知識を持っていなかったこと及び概念数が膨大であるために，DODDLE-OWL が提供する階層洗練戦略が示唆する概念変動箇所が千カ所以上におよび確認が困難であったことから，概念階層の洗練は行っていない．

5.2. 実験方法

5.1 節で述べた手順で構築したロケット運用オントロジーを GXFinder に搭載し，検索実験を行った．検索は，ロケット運用に詳しい専門家が行った．検索コンテクストを仮定せずに検索対象となる文書を専門家が思い浮かべることは困難であるため，領域オントロジーの概念階層を専門家に見てもらいながら検索対象となる文書を想定してもらった．検索対象となる文書を専門家が決定後，はじめにキーワード検索を行い，次に検索結果数が多い場合には絞込検索を，少ない場合には拡大検索を行った．キーワード検索と領域オントロジーを用いた検索の結果上位 10 件および 20 件について適合率，再現率，F 値を用いて評価を行った．

5.3. 実験結果

本節では実験 1 及び 2 の結果を示す．実験 1 は，「発射管制卓」に関する文書を検索したいが，「発射」が思い浮かばず，「管制卓」で検索を行うという想定で行った．実験 2 は，「ターミナルカウントダウン

シーケンス」に関する文書を検索したいが，「ターミナル」が思い浮かばず，「カウントダウンシーケンス」で検索を行うという想定で行った．表 1 および表 2 に実験 1 及び 2 の検索結果を示す．検索結果はキーワード検索と絞込検索における検索結果上位 10 件及び 20 件の適合率，再現率，F 値である．

実験 1 及び 2 の結果より，絞込検索はキーワード検索に比べ適合率，再現率，F 値のすべてにおいて良い結果を得ることができた．絞込検索では，領域オントロジーの階層構造を利用して，より詳しいキーワードに絞り込んで検索することにより，ユーザが検索キーワードを明確に思い浮かべることができない場合でも，目的の文書が検索可能であると考えられる．拡大検索については，専門家が全文書を詳細に把握していることもあり，適切に機能する例題が得られなかった．

表 1　実験 1 の検索結果

	再現率	適合率	F 値
キーワード検索 上位 10 件	0.250	0.100	0.143
絞込検索 上位 10 件	0.750	0.300	0.429
キーワード検索 上位 20 件	0.750	0.150	0.250
絞込検索 上位 20 件	1.000	0.200	0.333

表 2　実験 2 の検索結果

	再現率	適合率	F 値
キーワード検索 上位 10 件	0.136	0.300	0.188
絞込検索 上位 10 件	0.364	0.800	0.500
キーワード検索 上位 20 件	0.136	0.150	0.143
絞込検索 上位 20 件	0.500	0.550	0.524

5.4. 考察

本ケーススタディでは，拡大検索が適切に機能しなかった．原因として，キーワードに対応する概念の兄弟概念が非常に多く，拡大検索を行うと幅広くキーワードが展開されてしまい，検索ヒット数が膨大になったことが考えられる．兄弟概念数を減らす方向で領域オントロジーを洗練することにより，拡大検索が適切に機能すると考えられる．兄弟概念数を減らす方法の一つとして，DODDLE-OWL では語頭による複合語の階層構築を行っている．語尾のみによる複合語の階層構築に比べて，兄弟概念数を減らすことが可能だが不十分である．より兄弟概念数を減らすためには，構築中の領域オントロジーだけでなく，EDR 電子化辞書中の概念階層も考慮して，複合語の概念階層の洗練を行う方法が考えられる．

6. おわりに

本研究では，オントロジー搭載型検索システムおよび日本語オントロジーを半自動で構築可能なツールである DODDLE-OWL の提案を行った．ロケット運用分野におけるケーススタディを通して，DODDLE-OWL を用いることにより, 初期領域オントロジーを半自動で構築できることを示した. また, 領域オントロジーを用いた絞込検索が，キーワード検索よりも精度が高いことを示した．拡大検索については，初期概念階層の兄弟概念数が膨大なため，検索結果数も膨大となり，適切に機能しなかった．今後は，拡大検索が適切に機能するように領域オントロジーの洗練に取り組む予定である．また，検索ログを領域オントロジーの洗練に用いる方法についても検討していく予定である．

参考文献

[1] T.Morita, N.Fukuta, N.Izumi and T.Yamaguchi: DODDLE-OWL: A Domain Ontology Construction Tool with OWL, ASWC2006, LNCS4185, pp.537-551, 2006.

[2] S.Koide, M.Kawamura, T.Morita, T.Yamaguchi and H.Takeda: Semantic Search: An Implementation, Deployments, and Lessons Learned, ASWC2006 Workshop on Web Search Technology, 2006

[3] Sen, http://ultimania.org/sen/

[4] 日本電子化辞書研究所, EDR 電子化辞書(第 2 版)仕様説明書, TR2-006(改), 2001.