2E1-1

オントロジーマッピングに有効な特徴の抽出

Effective Features for Judging Ontology Mapping

市瀬 龍太郎[*1]
Ryutaro Ichise

[*1]国立情報学研究所
National Institute of Informatics

This paper presents an analysis of effective features for judging ontology mapping. We defined several concept similarity measures for judging and analyzed them by discriminant analysis. The results show that twenty two proposed measures out of forty eight are effective for judging ontology mapping and the effective measures include wide variations.

1. はじめに

現在，ホテルの予約情報，航空券の予約情報など，さまざまな情報が Web 上で入手可能である．しかし，現在の Web システムでは，それらの情報を人手で収集し，各々を処理しなければならない問題点がある．このような手間を解決するために，Web 上の情報を連携させるセマンティック Web 技術が注目を集めている．セマンティック Web を使って様々なデータを連携させるには，それぞれのデータがどのような意味を持つのかをオントロジーで付与させ，そのオントロジーを連携させる必要がある．そのために，異なるオントロジー間の対応関係を自動的に導出するオントロジーマッピング技術 [Euzenat 07] の開発が盛んに行われてきている．

オントロジーマッピングとは，異なるオントロジーがあった時に，それらの間にどのような対応関係があるのかを発見する問題である．例えば，あるオントロジーの中に，宿泊施設という概念があったとする．また，別のオントロジーの中に，宿という概念があったとする．その時に，宿泊施設と宿という記述が異なるものであっても，両者が記述している意味は一致していると判断できれば，両方のオントロジーで別々に記述されているホテルや旅館などから，宿泊する場所を選び出すことができる．このように，異なるオントロジーで記述されている概念などの対応関係を発見することをオントロジーマッピングと呼ぶ．

オントロジーマッピングは，[市瀬 07b] で述べられているように，文字列の類似性を利用した手法や，オントロジーのグラフを利用した手法などさまざまな解決アプローチが取られている．例えば，Aumueller [Aumueller 05] らは，対応関係をさまざまな手法で計測し，対応関係を決める COMA++というシステムの開発をしている．一方，市瀬は，オントロジーマッピング問題を，概念間の類似度を属性として利用することで，対応か非対応かを決定する機械学習問題として定式化している [市瀬 07a]．この研究により，従来から研究されてきた様々なオントロジーマッピングに使われる手法を類似度として利用することで，一つの枠組みに統合することが可能となった．しかし，これらの類似度の尺度に対して，どのようなものが有効であるかの解析がこれまでなされていなかった．そこで，本研究では，オントロジーマッピングに使われる様々な類似度の尺度に対して，解析を施すことにで，判別に有効な特徴についての議論を行う．

連絡先: 市瀬 龍太郎，国立情報学研究所情報学プリンシプル研究系，〒 101-8430 東京都千代田区一ツ橋 2-1-2，Tel:03-4212-2000，Fax:03-3556-1916，E-mail:ichise@nii.ac.jp

2. 概念の類似性の尺度

これまでに，オントロジーマッピングに使われる概念の類似性に対して，多くの類似性の指標が提案されてきた．代表的なものとして，文字列に基づく指標，グラフに基づく指標，インスタンスに基づく指標，知識に基づく指標などが挙げられる [市瀬 07b]．文字列に基づく類似性の指標は，オントロジーマッピングにおいて，しばしば用いられるものであり，概念ラベルの文字列などを利用して類似性を計算する．グラフに基づく指標は，オントロジーの構造に着目して類似性を計算する．オントロジーは，木構造をしているため，２つの木構造のグラフに対して，類似性を計算するのである．そのようなものを使ったシステムとして，Similarity Flooding [Melnik 02] や S-Match [Giunchiglia 04] などがある．インスタンスに基づく指標は，概念対における共有インスタンスの分類の類似性に着目して，類似性の計算を行う．そのようなものを使ったシステムとして，HICAL [市瀬 02] などがある．知識に基づく指標では，WordNet [Fellbaum 98] や辞書などのリソースを利用することで，類似度を計算する．このように，たくさんの類似度の尺度があるが，これまでに，どの類似度の尺度がオントロジーマッピングの判定に有効な特徴であるのかの分析がなされてこなかった．そこで，本研究では，[市瀬 07a] で用いられた４種類の類似度の尺度,「語類似度」,「語リスト類似度」,「概念階層類似度」,「構造類似度」を利用して，これらの指標の有用性を調べる．以下，上記の４つの尺度を順番に説明する．

2.1 語類似度

ここでは，概念の類似度を測る基本的な指標として，文字列に基づく４種類の類似度の指標と，知識に基づく４種類の類似度の指標を語類似度として述べる．

文字列に基づく類似度は，文字列を使って計算する．ここでは，以下の４種類の類似度を用いる．

- プレフィックス
- サフィックス
- 編集距離
- n グラム

プレフィックスは，語の先頭の類似度を測る指標で，Eng と English のようなものに対して，効果的に類似度を計算できる．サフィックスは，逆に語の末尾の類似度を測る指標で，phone と telephone のようなものに対して，効果的に類似度を計算できる．編集距離は，文字列の置換，削除，挿入の回数に基づいて，類似度を計算する．n グラムは，n 個の文字毎に語を分割し，同じものの数を類似度とする．例えば，2 グラムを用いた時には，word という語は wo,or,rd の３つに分けられ，別の語を同様に分けたものとの共通の部分を使って類似度を計算する．本論文では，3 グラムを用いている．

同様に，知識に基づく指標も文字列に対して計算する．本研究では，WordNet を知識リソースとして用いた．WordNet を使った類似度は多く提案されているが，本研究では，以下の４種類の指標を用いる．

- 同義語 (synset)
- Wu & Palmer
- 説明 (description)
- Lin

同義語 (synset) は，WordNet の同義語のパスの長さを利用した類似度の指標である．WordNet は，同義語 (synset) の情報が含まれているため，異なった語のペアに対して，最短のパス長を計算することができる．この類似度の指標は，このパス長を類似度として利用する．Wu & Palmer は，深さと最小共通上位概念 (LCS: least common superconcept) を用いて，下記の式に従って類似度を計算する [Wu 94].

$$similarity(W_1, W_2) = \frac{2 \times depth(LCS)}{depth(W_1) + depth(W_2)}$$

W_1 と W_2 は，概念のラベルを表し，$depth$ は，WordNet におけるその語の深さを表し，LCS は，W_1 と W_2 の最小共通上位概念を表す．説明 (description) は，WordNet におけるその語の説明を用いて類似度を計算する．各々の語の説明に共通する語の長さの２乗を使って類似度を計算する．最後の Lin [Lin 98] は，Wu & Palmer の式と同様であるが，深さの代わりに情報量を用いる．

2.2 語リスト類似度

次に，前節で述べた語類似度を語リストの類似度に拡張する．前節の語類似度は，語の類似度を測るための指標であるため，「Food_Wine」のような語のリストに対して類似度の計測ができない．しかし，このようなものは，概念のラベルとしてしばしば用いられる．もし，このような語をハイフンや下線で分割すると，語リストを得ることができる．そこで，この節では，このような語リストについて，最大語類似度と語編集距離の２種類の類似度を定義する．

最大語類似度では，２つの語リストの中の任意の語の組合せに対する語類似度の中で最大のものをその語リストの類似度とする．本論文では，前節で８種類の語類似度を定義した．従って，最大語類似度でも，８種類の異なる語リスト類似度が得られることになる．

語編集距離は，編集距離を文字列から語に拡張した類似度の指標である．例えば，{Pyramid} と {Pyramid, Theory} の２つの語リストの間の類似度を測ることを考える．ここで，語を編集距離計算の際の一つの文字列のように考えると，この語リストに対しても，編集距離を計算することが可能となる．この場合には，Pyramind が同じで，Theory の部分が異なるため，語編集距離は１と計算できる．しかし，{Social, Science} と {Social, Sci} のような場合には，Science と Sci を同じものであると判定するか否かの問題が生ずる．もし，同じと判定すると，語編集距離は０となるが，異なると判定すると１となる．文字列の場合には，同じものであるか否かは，容易に判定できるが，語の場合には，これが難しい．そこで，前節で述べた語類似度を再び用いることにする．ある閾値を用いれば，語類似度によりその語が同じか否かを判定することができる．例えば，プレフィックスを用いると，これらの例は同じ語であると判断できるが，同義語を用いた場合には，sci という語が WordNet に存在しないため，同じ語とは判定できない．その結果，プレフィックスを語類似度として用いた時には０，同義語を用いた場合には１と語編集距離を計算することが可能となる．使う語類似度に応じて，語編集距離では，８種類の異なる語リスト類似度が得られることになる．

以上の議論より，最大語類似度で８種類，語編集距離で８種類の合計１６種類の語リスト類似度が得られることになる．

2.3 概念階層類似度

この節では，オントロジーの概念階層の類似度について述べる．概念階層類似度では，オントロジーの概念階層のパスを用いて，類似度を計算する．表 1 の例を用いて説明しよう．ここでは，概念として，オントロジー A の Social_Sci とオントロジー B の Social_Science を対象と考える．この時，それぞれの概念は最上位の概念から，表 1 のパスの位置で表されるとする．その時，概念階層類似度を計算するために，パスをパスリストに分割する．すると，パスリスト中の概念ラベルを一つの文字列と見なせば編集距離を用いて，類似度が計算できる．ここで，語のリストを編集距離で計算する時と同じ問題が生ずる．すなわち，Social_Sci と Social_Science を同じものと見なすか否かである．ここで，語リスト類似度を語類似度を用いて計算した時と同様に，概念階層類似度を語リスト類似度を用いて計算する．つまり，表 1 の語リストのようにパスリストを分割し，それぞれの語リストに対して，語リスト距離を計算して，閾値により同じものと見なすか否かを決定する．その結果，１６種類の語リスト類似度があるため，１６種類の概念階層類似度が得られることになる．

2.4 構造類似度

この節では，構造に対する類似度を決める．前節で概念階層に対する類似度を決めたが，これだけではグラフ的な構造に対する類似度を取り扱うことができない．そこで，対象概念の近傍の概念となる親概念を使って，構造の類似度を測ることにする．親概念のラベルに対して，類似度を計算するには，語リスト類似度を用いることができる．従って，構造類似度として，16 種類の類似度を定義できる．

3. オントロジーマッピングの判定に有効な特徴

本研究では，前章で述べた 16 種類の語リスト類似度，16 種類の概念階層類似度，16 種類の構造類似度の計 48 種類の類似度の指標に対して，どの指標がオントロジーマッピングの判定に有効であるかの分析を行った．

類似度の指標を解析するために，本研究では，オントロジーマッピングの性能評価用データとして公開されているインターネットディレクトリのデータを用いた．このデータは，Ontology Alignment Evaluation Initiative(OAEI) [OAE 08] が 2005 年

表 1: 概念階層類似度を計算する時の例

	パス	パスリスト	語リスト
オントロジー A	Top / Social_Sci	{Top, Social_Sci}	{Top}, {Social, Sci}
オントロジー B	Top / Social_Science	{Top, Social_Science}	{Top}, {Social, Science}

に性能評価ワークショップのために提供したものである．このデータは，実際に使われている 3 つのインターネットディレクトリから，単純な概念階層を取り出し，全部で 2265 個のペアに対して，人手でマッピングを付けたものである．このデータには，いくつかのエラーが含まれているため，それらを取り除いた 2193 個のデータを利用した．このデータには，人手で付けられた正しいマッピング（正例）が含まれているが，正しくないマッピング（負例）が含まれていない．そこで，正例になっている 2 つの概念対を取り出し，一方の概念を固定し，他方の概念をその概念が含まれるオントロジーの中の別の概念に置き換えることで，負例の概念対を生成した．この負例は，厳密な意味において，全てが完全に間違えたマッピングであるとは言えないが，正例を人手で付けているため，それよりは，劣ったマッピングであるという点で，負例として妥当であると言える．

この実験データを用いて，与えられた 48 種類の類似度の指標に対して，判別分析を行い，特徴の寄与度についての解析をした．解析の際には，もっとも有効な説明変数（指標）を判別式に順次取り入れていく変数増加法を用いた．また，変数を増加させる際の有意水準として，5%を利用した．

解析の結果，48 種類の類似性尺度の中から，表 2 に示した 22 種類の類似性の尺度が抽出された．これらの尺度は，オントロジーマッピングの判定に有効な特徴と言うことができる．表の左にある比較対象とは，異なるオントロジーの何を比較して抽出された特徴かを示している．ここでは，概念，概念階層，構造の 3 つがあり，それぞれ，2.2 節で定義した概念同士の比較，2.3 節で定義した概念階層同士の比較，2.3 節で定義した構造の比較を表している．表 2 中央の語リスト手法とは，比較にどのタイプの語リスト類似度を使ったのかを示している．ここには，最大の語の類似度を用いた最大語類似度と編集距離を語に拡張した語編集距離の 2 種類がある．表 2 右のベース手法とは，基本となる語同士の比較にどの手法を用いているかを表している．これには，2.1 節で定義したプレフィックス，サフィックス，編集距離，n グラム，同義語，Wu & Palmer，説明，Lin の 8 種類が入ることとなる．

まず，比較対象に関して表 2 を見ると，概念が 7 個，概念階層が 8 個，構造が 7 個とバランスよく分散されていることが分かる．従来のオントロジーマッピングシステムにおいては，概念同士の類似度の比較が判定に有効であるとして，多く用いられる傾向にあるが，この結果より，実際には，概念同士の類似度だけでは十分でなく，概念階層や構造も合わせて比較をしないとオントロジーマッピングに有効な特徴をとらえることができないことが分かる．一方，判別に有効な特徴を上位から順に見ると，概念階層が上位の方に並んでいる．従って，大雑把なマッピングを判定する時には，概念階層が大きな役割を果たしていると言えるであろう．次に，語リスト手法に関して表 2 を見ると，最大語類似度が 9 個，語編集距離が 13 個となっており，数的には語編集距離の方が少し多いと言える．しかし，最大語類似度は，語編集距離よりも上位に多く出現している．このことより，最大語類似度は，大雑把な分類の時には，重要な役割を果たしているが，詳細な分類を行うには，語編集距離が不可欠であると言えるであろう．次に，ベース手法に関して表 2 を見ると，プレフィックス 3 個，サフィックス 1 個，編集距離 4 個，n グラム 1 個，同義語 2 個，Wu & Palmer4 個，説明 4 個，Lin3 個となっている．これに関しても，全ての手法が出現しており，どの指標もオントロジーマッピングの判定に有効な特徴になっていることが分かる．この指標を 2 章で述べた文字列に基づく指標，知識に基づく指標という観点から見ると，前者が 9 個，後者が 13 個となる．これに関しても大きな差があるとは言い難いが，知識を利用したものの方が少し多い．一般的には，文字列に基づく指標が手軽なために多く使われているが，知識に基づく指標には，それ以上に有効な指標であると言えるであろう．

表 2: オントロジーマッピングの判定に有効な特徴

比較対象	語リスト手法	ベース手法
構造	最大語類似度	編集距離
概念	最大語類似度	編集距離
概念階層	語編集距離	Lin
概念階層	最大語類似度	編集距離
概念階層	語編集距離	説明
概念階層	最大語類似度	説明
概念階層	語編集距離	プレフィックス
概念階層	最大語類似度	Lin
概念階層	最大語類似度	同義語
構造	最大語類似度	Wu & Palmer
概念	語編集距離	N グラム
概念	最大語類似度	Wu & Palmer
構造	語編集距離	Lin
概念階層	語編集距離	Wu & Palmer
概念	語編集距離	Wu & Palmer
構造	語編集距離	説明
構造	語編集距離	サフィックス
構造	語編集距離	同義語
概念	最大語類似度	説明
概念	語編集距離	編集距離
概念	語編集距離	プレフィックス
構造	語編集距離	プレフィックス

全体として見ると，本研究で使われた 48 種類の指標のうち，22 個しか有効であるとは判定されなかったが，2 章で定義した全ての指標を偏りなく使用していた．このことは，オントロジーマッピングの判定に対して，決定的な特徴がないことを示しており，さまざまな特徴の組み合わせによって判定を行う必要性を示していると言える．

抽出された 22 個の特徴を用いてオントロジーマッピングを判別した結果は，図 1 のようになった．図中の Group1 と Group2 は，それぞれが正例 (対応)，負例 (非対応) を表しており，横軸が値，縦軸が割合を示している．また，正答率は，73.78%であった．このグラフより，判別する部分は，かなり近接しており，判別が難しい問題であることが分かる．また，線形分離可能な前提においては，73.78%しか分離ができない

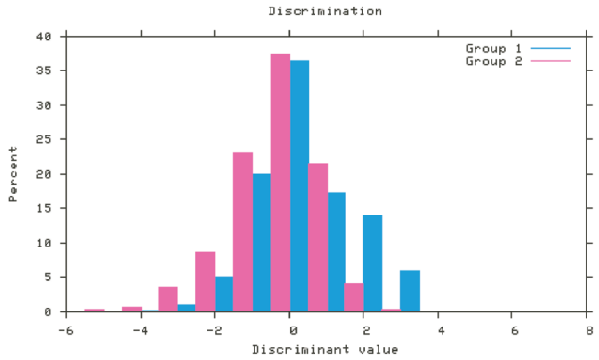

図 1: オントロジーマッピングの判別結果

ことより，本研究で取り上げたオントロジーマッピング問題では，非線形の学習手法を取り入れる必要性や，属性を強化する必要性があると考えられる．

4. おわりに

本研究では，オントロジーマッピングの判定に有効な特徴の抽出を試みた．そのために，オントロジーを比較するための様々な類似度を取り上げ，それらの指標がどの程度，判別に有効かの解析を行った．その結果，本論文で定義した 48 種類の指標のうち，22 種類の指標が抽出された．しかし，それらの指標は多岐に渡っており，オントロジーマッピングの判定には，さまざまな特徴を利用しなければならないことが示された．

今後の課題としては，まず，属性の強化が必要であると考えられる．現在，線形分離可能との前提で，73.78%しか正しい判定ができないため，判定に有効な属性をまだ加える必要があると考えられる．そのためには，Pedersen らが提案する語の類似性 [Pedersen 04] などを属性として新たに取り込むことが考えられる．一方，オントロジーマッピング問題は，線形分離可能な問題ではないととらえることも可能である．[市瀬 07a] では，そのような場合に使える学習器の代表である SVM を利用しているが，今後は，さらにいろいろなオントロジーマッピングのデータセットを検証することで，適切な学習手法を同定し，正答率を上げる手法を探っていく必要性がある．

参考文献

[Aumueller 05] Aumueller, D., Do, H. H., Massmann, S., and Rahm, E.: Schema and ontology matching with COMA++, in Özcan, F. ed., *Proceedings of the ACM SIGMOD International Conference on Management of Data*, pp. 906–908, ACM (2005)

[Euzenat 07] Euzenat, J. and Shvaiko, P.: *Ontology Matching*, Springer (2007)

[Fellbaum 98] Fellbaum, C.: *Wordnet: An Electronic Lexical Database*, MIT Press (1998)

[Giunchiglia 04] Giunchiglia, F., Shvaiko, P., and Yatskevich, M.: S-Match: an Algorithm and an Implementation of Semantic Matching, in Bussler, C., Davies, J., Fensel, D., and Studer, R. eds., *Proceedings of the 1st European Semantic Web Symposium*, Vol. 3053 of *Lecture Notes in Computer Science*, pp. 61–75, Springer (2004)

[Lin 98] Lin, D.: An information-theoretic definition of similarity, in *Proceedings of the 15th International Conference on Machine Learning*, pp. 296–304, Morgan Kaufmann, San Francisco, CA (1998)

[Melnik 02] Melnik, S., Garcia-Molina, H., and Rahm, E.: Similarity Flooding: A Versatile Graph Matching Algorithm and its Application to Schema Matching, in *Proceedings of the 18th International Conference on Data Engineering*, San Jose, CA (2002)

[OAE 08] Ontology Alignment Evaluation Initiative, http://oaei.ontologymatching.org/ (2008)

[Pedersen 04] Pedersen, T., Patwardhan, S., and Michelizzi, J.: WordNet::Similarity - Measuring the Relatedness of Concepts, in *Proceedings of the 19th National Conference on Artificial Intelligence*, pp. 1024–1025 (2004)

[Wu 94] Wu, Z. and Palmer, M.: Verb semantics and lexical selection, in *Proc. of the 32nd Annual Meeting of the Association for Computational Linguistics*, pp. 133–138, New Mexico State University, Las Cruces, New Mexico (1994)

[市瀬 02] 市瀬 龍太郎, 武田 英明, 本位田 真一：階層的知識間の調整規則の学習, 人工知能学会論文誌, Vol. 17, No. 3, pp. 230–238 (2002)

[市瀬 07a] 市瀬 龍太郎：機械学習問題としてのオントロジーマッピング, 人工知能学会研究会資料, Vol. SIG-FPAI-A603, pp. 59–64 (2007)

[市瀬 07b] 市瀬 龍太郎：情報の意味的な統合とオントロジー写像, 人工知能学会誌, Vol. 22, No. 6, pp. 818–825 (2007)