知識ベースに対するプロパティ指向のファセット検索システムに関する研究

阿曽　太郎†　　天笠　俊之††　　北川　博之††

† 筑波大学システム情報工学研究科　〒 305–8573 茨城県つくば市天王台 1-1-1
†† 筑波大学計算科学研究センター　〒 305–8573 茨城県つくば市天王台 1-1-1
E-mail: †aso@kde.cs.tsukuba.ac.jp, ††{amagasa,kitagawa}@cs.tsukuba.ac.jp

あらまし　知識ベースとは様々な知識が蓄積されたデータベースである．本研究では，その中でも RDF で記述された知識ベースに焦点を当てる．一般に，RDF による知識ベースのデータ構造は複雑であるため，専門知識を持たないユーザが簡単に検索を行うには，ファセット検索が有効であることが知られている．ファセットとはデータの切り口であり，ユーザはファセットの選択・解除を繰り返すことで対話的な検索を実行できる．本論文では，RDF の述語（プロパティ）について，主語と述語との構造的な関係に着目してクラスタリング することで，プロパティに関するファセットを生成することを提案する．

キーワード　知識ベース，RDF，ファセット検索，クラスタリング

1　はじめに

知識ベースとは，様々な知識が蓄積されたデータベースである．代表的な知識ベースには，Wikipedia の情報を基にした DBpedia や Wikidata，YAGO などがある．人間や機械は知識ベースを使うことで，質問に答えたり，新たな知識を発見することができる．

知識ベースの記述には，Resource Description Framework（RDF）が用いられる．RDF とは，リソースに関する情報を記述する方法である．RDF では，Universal Resource Identifier（URI）で識別されるものすべてをリソースとして扱う．世の中のあらゆるエンティティは URI を付けることで，リソースとして記述することができる．あるリソースについての 1 つの情報は，主語（Subject），述語（Predicate），目的語（Object）から構成される 3 つ組（トリプル）のグラフ構造で記述される (図 1)．主語は情報を記述される対象のリソースを示し，述語は主語に関する情報のプロパティを定義する．そして，目的語は述語の対象である．主語と述語は URI で記述し，目的語は URI もしくは数値や文字列などのリテラルで記述する．

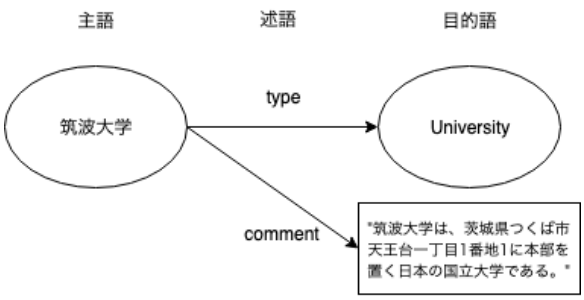

図 1　RDF の例

RDF で記述された知識ベースに対して検索を行うにはいくつかの方法がある．主要なものの 1 つは問合せ言語 SPARQL を用いた検索である．SPARQL の文法に従って，トリプルの条件を指定することで，情報を取り出すことができる．しかし，一般ユーザにとって，SPARQL 検索を行うハードルは高い．なぜなら，まず，SPARQL の文法を理解し習得する必要がある．そして，検索対象の知識ベースで定義されているプロパティやエンティティについて理解する必要があるためである．知識ベースは様々な種類のプロパティやエンティティが存在する複雑なグラフ構造になっているため，特に後者の理解は難しい．もう 1 つの方法として，キーワード検索がある．キーワード検索は，キーワードを入力することで検索が行えるため，前提知識が不要な検索手法である．結果には，キーワードに関するエンティティのランキングが返却される．しかし，様々な種類のエンティティが混在しているため，ユーザは欲しい情報を判断しづらい．また，ユーザはどのような種類のエンティティがあるのか把握していない場合や，そもそもどのような種類のエンティティが欲しいのかわかっていない場合もある．したがって，前提知識が不要という点で，キーワード検索は有効な手段だが，情報を整理して提示する必要があると考える．

こうした課題を解決する検索方法としてファセット検索がある．ファセット検索では，検索対象のエンティティを様々な切り口（ファセット）で絞り込む．ユーザは結果を確認し，ファセットを切り替えたり，組み合わせたりすることで，意図する結果を得るまで，対話的に検索を行うことができる．したがって，知識ベースのプロパティやエンティティの種類などの知識を持たない場合でも，検索を容易に実行できる．これまでに提案されてきた RDF の知識ベースに対するファセット検索システムの多くは，エンティティが持つプロパティをファセットとして利用してきた．しかし，プロパティそれ自体も数多くの種類が存在しているため，必要となるプロパティ（ファセット）を見つけ出すことは簡単ではない．この課題を解決するには，

数多くあるプロパティから興味のあるプロパティを見つけ易くすることが必要である．そのために，プロパティをその主語と目的語との関係性によってクラスタリング した結果をプロパティに関するファセットとして利用することを提案する．そして，プロパティに関するファセットを利用して，関係するエンティティ集合（トリプル集合）を検索できる，プロパティ指向の新しいファセット検索システムを提案する．

本稿では，プロパティ指向のファセット検索システムの全体概要，プロトタイプシステムのインターフェース，機能要素であるファセットに関してプロパティのクラスタリング について報告する．

2 前提知識

本節では，前提知識として，ファセット検索について説明する．

2.1 ファセット検索

ファセット検索とは，探索的検索における手法の 1 つである．その特徴は，検索対象のエンティティ集合を，様々な切り口（ファセット）によって絞り込むことで，意図するエンティティを発見しようとするところにある．例えば，E-Commerce の Amazon の商品検索サイトもファセット検索である．基本的なインターフェースとして図 2 を例にあげる．ユーザはキーワード検索によって，Extension に初期の検索結果を，Transition Marker に初期のファセットを得る．その状態から，ファセットの選択や，選択したファセットの解除を対話的に繰り返して，検索結果を洗練させていく．

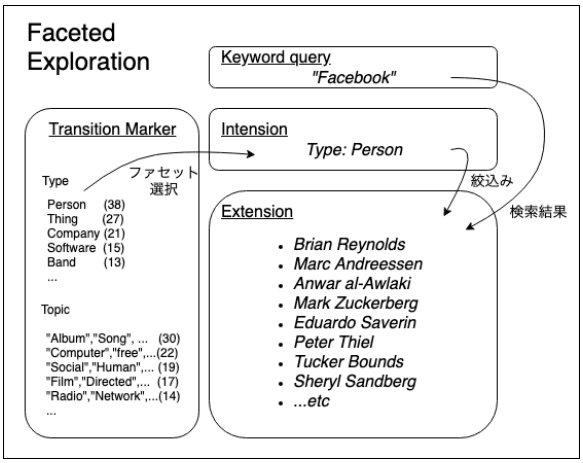

図 2　ファセット検索のインターフェースイメージ

3 関連研究

3.1 RDF の検索結果に対するランキング

RDF の検索結果の有用性を向上させることを目的に，Ichinose ら [4] は DBpedia に対して PageRank を用いてエンティティを評価する手法を提案している．主語と目的語をノードとし，述語をエッジとして PageRank を計算し，エンティティの重要度を評価している．この研究の実験では，検索対象のエンティティが既知，あるいは SPARQL クエリにおいてトリプルの条件が設定されており，対象ユーザは SPARQL や対象データの DBpedia に関する知識を持っていることを前提としている．本研究では，エンティティの評価には PageRank を用いるが，ユーザは SPARQL や知識ベースに関する知識を有していないケースを想定しているため，前提条件が異なる．

3.2 RDF に対するキーワード検索

RDF に対するキーワード検索を行う研究として，奥村ら [6] の ObjectRank と適合フィードバックを用いた研究がある．検索対象となるエンティティはリテラルを目的語に持つトリプルを持つことを前提にして，エンティティをリテラルを含めたドキュメントとみなしている．本研究でもこの考え方に則る．しかし，検索対象となるエンティティの種類がシステム構築者などによってあらかじめ設定されることを前提とした手法であるため，知識ベースに含まれる多様なエンティティの全てに応える方法ではない．また，ユーザの検索意図に応える方法として，適合フィードバックを採用している点も，ファセット検索とはアプローチが異なる．

3.3 知識ベースに対するファセット検索

知識ベースに対するファセット検索の研究は数多く行われている．本稿の対象データセットである DBpedia に関しては，Brunk ら [3] が，DBpedia のオントロジーを利用して，階層的なタイプ情報をファセットとして選ばせる tFacet を提案している．Arenas ら [1] は，知識ベースの YAGO に対してファセット検索を行う SemFacet を提案している．SemFacet では，ファセットに RDF における目的語やプロパティを使用し，検索対象にエンティティ（URI）を設定している．エンティティに関する情報が疎である場合があることが課題とされるが，OWL2 のオントロジーを用いて推論することによって解決することを提案している．Papadakos ら [7] は，ファセットをランキングする Hippalus というシステムを提案している．Hippalus では，ユーザが検索プロセスの中でファセットを評価し，そのファセットの評価に基づいてファセットをランキングして返す．これにより，ユーザの好みに合わせたファセット検索が行えるとしている．Bast ら [2] は，知識ベースの 1 つであった Freebase のファセット検索を提案している．主な特徴は，ファセット検索の利便性を向上させるために，データセットそのものを編集したことにある．具体的には，データセットに含まれる冗長なエンティティやプロパティを削除や統一，タイプのタキソノミー（分類）の編集などを行なっている．Wikidata に関しては，Moreno-Vega ら [5] が大規模な知識ベースに対するクエリの高速化を目指した GraFa を提案している．

上記の研究では，既存のプロパティをエンティティのファセットとして利用し，エンティティを探索できるようにすることを目的としている．一方で，本研究では，エンティティ間の関係性および関係付けられているエンティティの探索を重視する．そのために，エンティティ間の関係性を定義するプロパティに関するファセットを生成し，プロパティの探索を従来より容易

にすることを目指している点が新しい．

4 提 案 手 法

知識ベースに対するプロパティ指向のファセット検索システムとして，図 3 のシステム概要を提案する．プロパティ指向のファセット検索システムの目的は，興味に関してキーワード検索して得たトリプルの集合に対して，エンティティ間の関係性を示すプロパティに関するファセットを整備することで，関係性に基づいてエンティティの集合（トリプルの集合）を検索できるようにすることである．ユーザの操作とシステムで実行される処理は次の通りである．

（1） ユーザはキーワード検索を実行する．システムは，キーワードをエンティティのドキュメントに含むエンティティについて，それらが主語または目的語に位置付けられているトリプル集合を主語あるいは目的語のランク値降順で返却する．また，返却したトリプル集合に含まれる主語，述語，目的語についてファセットを提示する．

（2） ユーザはファセットから探索したいファセットキーを選択する．システムは，選択されたファセットキーに対応するトリプル集合を返却する．

本システムのポイントは，プロパティに関するファセットのような元の知識ベースには存在しないファセットの生成を可能とするために，知識ベースから独立して関係データベースを設計することにある．以降で，RDF データベース，エンティティデータベース，トリプルデータベース，ファセットデータベースを説明し，最後にプロトタイプシステムである “ProFacet” のユーザーインターフェースについて説明する．

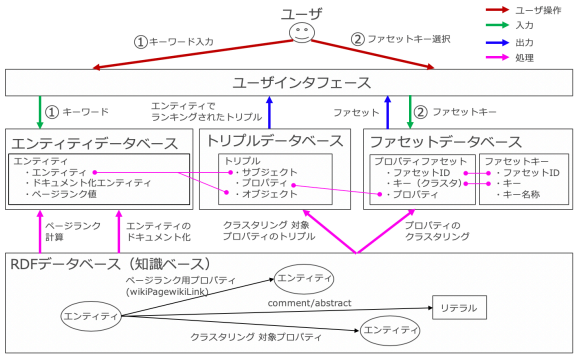

図 3 システム概要

4.1 RDF データベース

RDF データベースは，キーワード検索とファセット検索に必要となる知識ベースのデータを格納する．本稿では，検索結果として返却される RDF データと検索対象となるドキュメント化されたエンティティを生成するための RDF データの 2 種類を使用する．また，検索結果として返却される RDF データは，エンティティのランク値の計算とファセット生成のためのクラスタリング においても使用される．

4.2 エンティティデータベース

エンティティデータベースは，エンティティを主キーとして，検索対象のドキュメント化されたエンティティ，ランク値をタプルとしたテーブルを持つ．エンティティのランク値は，検索結果として返却される RDF データの主語と目的語のエンティティに対して，PageRank のアルゴリズムによって計算する．この時，PageRank は，複数種類のエッジには対応しないため，元の RDF データが持つ複数種類のプロパティの区別は行なっていない．そして，ドキュメント化されたエンティティを生成するための RDF データに対して，主語のエンティティと目的語のリテラルを 1 つのドキュメントとする処理を行う．これら 2 つの処理の結果を合わせて，エンティティ，ドキュメント化エンティティ，ランク値をタプルとしたエンティティテーブルを作成する．

4.3 トリプルデータベース

クラスタリング 対象のプロパティが述語であるトリプルをタプルとしたテーブルを持つ．インターフェースの検索結果には，このテーブルのタプルが表示される．

4.4 ファセットデータベース

ファセットデータベースは，エンティティに関するファセットやプロパティに関するファセットのテーブルと，各ファセットのキー名称を管理するテーブル “ファセットキー” を持つ．図 3 では，プロパティに関するファセットを例示している．ファセットのテーブルは，ファセットの種類を示す番号，ファセットの内容を示すキー，エンティティやプロパティの URI をタプルとして持つ．ファセットキーは，ファセットの種類を示す番号，ファセットキー，キーの内容を示すラベルをタプルとして持つ．ファセットの種類を示す番号は，ファセットの種類の拡張に対応するためである．また，ファセットキーは，エンティティやプロパティのクラスタ番号に該当する．本稿では，プロパティファセットを群平均法による階層型クラスタリング によって生成した．プロパティ間の距離は，Jaccard 係数を変換した Jaccard 距離である．Jaccard 係数は 2 つの集合に含まれる要素のうち，共通要素が占める割合を示す．ここでは，各プロパティの主語に関する Jaccard 係数と目的語に関する Jaccard 係数の平均値をプロパティ間の Jaccard 係数とした．この手法を適用した理由として，プロパティはエンティティとエンティティ（またはリテラル）の関係性として機能するため，主語や目的語を共有するという観点でプロパティを整理することができると考えたからである．別の方法として，RDF スキーマに基づくプロパティの階層構造を示すプロパティである “subPropertyOf” の利用や，プロパティの主語や目的語に期待されるエンティティのクラスを示すプロパティである “domain”，“range” なども考えられるが，あくまでも期待値であるため，実際のデータに基づいてクラスタリング を行うことが有効と考えた．

4.5 ユーザーインターフェース

プロパティ指向のファセット検索システム “ProFacet” のプ

ロトタイプを実装した．そのユーザーインタフェースを図 4 に示す．A のように，ユーザは検索キーワードや URI を入力し，主語あるいは目的語のエンティティに対して検索を実行する．検索結果は D のように主語，述語，目的語のトリプルのタプルで表示される．また，B は Transition Marker として，D の検索結果に対応するファセットを表示する．プロトタイプシステムでは，主語と目的語のエンティティのファセット（Subject Type, Object Type）に各エンティティのクラス情報を，述語のプロパティのファセット（Predicate Type）には上述したクラスタリング 結果を整備した．C は，Intension として検索の状態を示している．

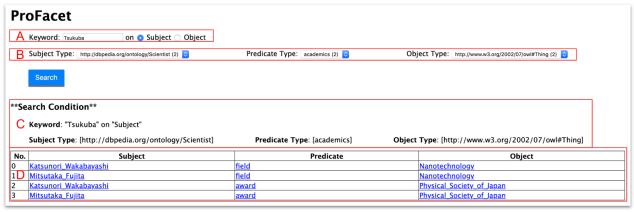

図 4　ユーザーインターフェース

5　予 備 実 験

プロパティ指向のファセット検索システムの機能要素となるプロパティに関するファセットの生成とその結果について目視による初期的な確認を行うことを目的として，次の 2 種類のプロパティのクラスタリング 結果についての確認を行なった．

（1）　上位プロパティによるクラスタリング

（2）　階層型クラスタリング によるクラスタリング

また，プロトタイプシステムを用いて，階層型クラスタリングによって生成したプロパティに関するファセットを用いた検索結果の確認を行った．

5.1　クラスタリング

5.1.1　上位プロパティによるクラスタリング

知識ベースの語彙の体系に基づいて，プロパティをクラスタリング する．プロパティは，RDF スキーマと呼ばれる語彙を定義する基本的な仕組みによって，そのプロパティの性質が定義されている．その中で，上位プロパティを定義するために使用される rdfs:subPropertyOf というプロパティを利用して，各プロパティを上位プロパティでクラスタリング した．クラスタの名称は上位プロパティの名称である．

5.1.2　階層型クラスタリング によるクラスタリング

実際のデータの関係性に基づいて，プロパティをクラスタリング する．クラスタリング の手法は群平均法による階層型クラスタリング を用いた．クラスタリング に用いたプロパティの類似度には，プロパティの主語に対する Jaccard 係数と目的語に対する Jaccard 係数の平均値を計算し，Jaccard 距離に変換した値を用いた．したがって，各クラスタの要素は，主語と目的語の重複度合いが類似するプロパティとなっている．クラスタの分割はデンドログラムを確認し，分割の閾値を 0.990 として実行した．クラスタの名称は，クラスタの要素数が 1 つしかない場合は，その要素の名称をそのまま用いた．また，要素数が複数ある場合は，各要素の内容を確認して便宜的に付けた．

5.2　実 験 環 境

提案手法を Python 3.7.3 で実装し，Intel(R) Core^{TM}i7-7700 3.60 GHz CPU, 32 GB RAM を搭載した Ubuntu 18.04.3 LTS で実験を行った．

5.3　データセット

本実験では，DBpedia 2016-10[1] の Instance Types において，クラスが University、Company、Scientist、Politician、Astronaut であるエンティティを抽出し，それらのエンティティを記述するトリプルを Mappingbased Objects のデータセットから抽出し，データセットとした．データセットの統計情報を図 5 に示す．

	Astronaut	*Company*	*Politician*	*Scientist*	*University*	*Total*
Number of Subjects	635	54570	16986	23423	20214	115828
Number of Predicates	11	29	54	31	28	101
Number of Objects	1536	117567	27829	51067	39461	219968
Number of Triples	5157	349860	78733	172382	113587	719719

図 5　実験に用いたデータセットの統計情報

5.4　上位プロパティによるクラスタリング 結果

上位プロパティを使うと，101 個のプロパティを 12 個のクラスタにまとめられた（図 6）．結果から，上位プロパティには名前空間の異なるプロパティも定義されていることがわかる．

また，図 7 は，最もサイズの大きなクラスタである “sameSettingAs” に含まれるプロパティの一覧である．“sameSettingAs” のプロパティは，rdfs:comment（リソースの説明文を記述するプロパティ）によると，“A relation between two entities participating in a same Situation; e.g., ’Our company provides an antivenom service’ (the situation is the service, the two entities are the company and the antivenom).” とあるが，直感的には，“sameSettingAs” と図 7 のプロパティの関係性を理解することは難しいと思われる．

Clusters	Number of Properties
http://www.ontologydesignpatterns.org/ont/dul/DUL.owl#sameSettingAs	32
http://www.ontologydesignpatterns.org/ont/dul/DUL.owl#coparticipatesWith	26
http://www.ontologydesignpatterns.org/ont/dul/DUL.owl#hasLocation	17
http://www.ontologydesignpatterns.org/ont/dul/DUL.owl#isMemberOf	7
http://www.ontologydesignpatterns.org/ont/dul/DUL.owl#isClassifiedBy	3
http://www.ontologydesignpatterns.org/ont/dul/DUL.owl#hasRole	3
http://dbpedia.org/ontology/location	2
http://www.ontologydesignpatterns.org/ont/dul/DUL.owl#isDescribedBy	2
http://www.ontologydesignpatterns.org/ont/dul/DUL.owl#isPartOf	2
http://www.ontologydesignpatterns.org/ont/dul/DUL.owl#isParticipantIn	2
http://www.ontologydesignpatterns.org/ont/dul/DUL.owl#hasSetting	1
http://www.ontologydesignpatterns.org/ont/dul/DUL.owl#conceptualizes	1

図 6　上位プロパティによるクラスタリング 結果

1 : https://wiki.dbpedia.org/downloads-2016-10

appointer	predecessor	monarch	runningMate
athletics	president	nominee	service
child	primeMinister	owner	spouse
citizenship	principal	owningCompany	subsidiary
education	provost	parent	successor
incumbent	rector	parentCompany	superintendent
influencedBy	relation	parentOrganisation	viceChancellor
keyPerson	relative	partner	vicePresident

図 7　“sameSettingAs” に属すプロパティの一覧

5.5　階層型クラスタリング によるクラスタリング 結果

階層型クラスタリング では，101 個のプロパティについて 23 個のクラスタを生成した（図 8）．クラスタに含まれる要素を確認すると，一定の領域に関係するエンティティに対して関係付けられるプロパティでまとまっていることが確認できた．図 9 は，University クラスタと Academics クラスタのプロパティの一覧である．大学という物理的・組織的なエンティティに関係付けられるプロパティと，学者や研究といったエンティティに関係付けられるプロパティに分けることができている．

clusters	*Number of properties*
company	20
academics	15
university	11
professionals	11
personalInformation	7
military	6
politics	4
governance	4
relationship	4
organization	3
soundRecording	2
principal	2

clusters	*Number of properties*
director	2
ethnicity	1
deathCause	1
appointer	1
language	1
dean	1
provost	1
vicePresident	1
officerInCharge	1
depiction	1
partner	1

図 8　階層型クラスタリング によるクラスタリング 結果

Properties in "University" cluster
affiliation
athletics
campus
chancellor
city
country
head
sport
state
viceChancellor
differentFrom

Properties in "Academics" cluster
academicAdvisor
almaMater
award
birthPlace
citizenship
deathPlace
doctoralAdvisor
doctoralStudent
field
influenced
influencedBy
knownFor
nationality
notableStudent
residence

図 9　“University” クラスタと “Academics” クラスタのプロパティ一覧

5.6　プロパティに関するファセットを用いた検索

主語のエンティティに対して，“Ivy League” でキーワード検索を行うと仮定し，検索結果を確認する．

5.6.1　キーワード検索結果

キーワード検索の結果，主語のエンティティのドキュメントに “Ivy League” を含むトリプルを 359 個取得した．内訳は，主語のエンティティの種類数が 39 個，プロパティの種類数が 26 個，目的語のエンティティの種類数が 137 個である．

キーワード検索結果に対応するプロパティに関するファセット（Predicate Type）は，図 10 の赤枠のように生成されている．また，主語のファセット（Subject Type）には 2 つのクラスが，目的語のファセット（Object Type）には，21 個のクラスがファセットキーとして生成されていることを確認した．この結果を踏まえて，ユーザは Predicate Type から University を選択し，大学に関係するようなエンティティ間の関係性について探索すると仮定する．

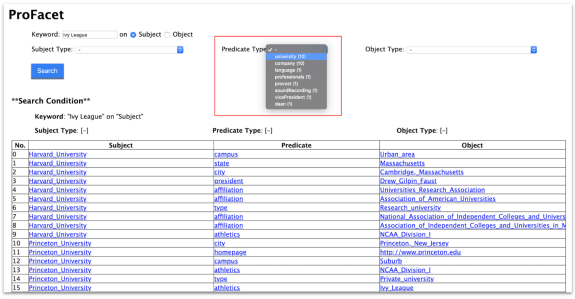

図 10　キーワード検索結果の一部と対応するプロパティファセット

5.6.2　ファセット検索結果

University を選択した結果，215 個のトリプルを取得した．内訳は，主語のエンティティの種類数が 30 個，プロパティの種類数が 10 個，目的語のエンティティの種類数が 89 個である．プロパティは大学に関係するようなプロパティのみになっていることが確認できる（図 11）．また，この時，主語のエンティティのクラスは 1 種類（University）となり，目的語のエンティティのクラスは 15 種類となっていることを確認した．このようにして，ユーザは，大学に関係するようなプロパティに絞り込んだ上で，University のクラスに属す主語のエンティティと他のクラスに属す目的語のエンティティとの関係性を探索することができる．

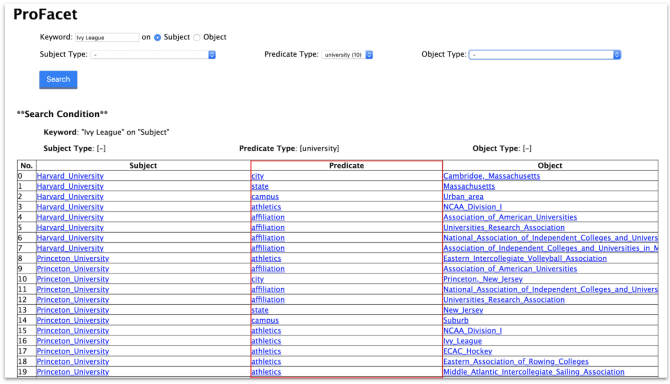

図 11　Predicate Type “University” のファセット検索結果の一部

6　まとめと今後の課題

本稿では，エンティティ間の関係性および関係付けられているエンティティ集合の探索を容易にするという目的に対して，プロパティ指向のファセット検索システムの概要を提案し，そ

の機能要素となるプロパティに関するファセット生成のためのクラスタリング に関する予備実験の結果と，提案システムのプロトタイプを用いた検索結果について報告した．予備実験では，プロパティの主語と目的語に関する Jaccard 係数を類似度とした階層型クラスタリング によって，一定程度関係する領域でプロパティをまとめられることを確認した．また，プロトタイプシステムを用いてプロパティに関するファセットが機能することを確認した．

今後の課題として，提案システムの有効性について既存のファセット検索システムと比較したユーザ調査を行う．

7 謝　　辞

本研究の一部は，SKY 株式会社との共同研究による．

文　　献

[1] Marcelo Arenas, Bernardo Cuenca Grau, Evgeny Kharlamov, Sarunas Marciuska, Dmitriy Zheleznyakov, and Ernesto Jimenez-Ruiz. Semfacet: Semantic faceted search over yago. In *Proceedings of the 23rd International Conference on World Wide Web*, WWW '14 Companion, pp. 123–126, New York, NY, USA, 2014. ACM.

[2] Hannah Bast, Florian Bäurle, Björn Buchhold, and Elmar Haußmann. Easy access to the freebase dataset. In *Proceedings of the 23rd International Conference on World Wide Web*, WWW '14 Companion, pp. 95–98, New York, NY, USA, 2014. ACM.

[3] Sören Brunk and Philipp Heim. tfacet: Hierarchical faceted exploration of semantic data using well-known interaction concepts. In *DCI@INTERACT*, 2011.

[4] Shiori Ichinose, Ichiro Kobayashi, Michiaki Iwazume, and Kouji Tanaka. Ranking the results of dbpedia retrieval with sparql query. In *JIST*, 2013.

[5] José Moreno-Vega and Aidan Hogan. Grafa: Scalable faceted browsing for rdf graphs. In Denny Vrandečić, Kalina Bontcheva, Mari Carmen Suárez-Figueroa, Valentina Presutti, Irene Celino, Marta Sabou, Lucie-Aimée Kaffee, and Elena Simperl, editors, *The Semantic Web – ISWC 2018*, pp. 301–317, Cham, 2018. Springer International Publishing.

[6] 奥村彩水, 天笠俊之, 北川博之. リンク構造解析を用いた linked open data に対するキーワード検索. In *DEIM Forum*, 2016.

[7] Panagiotis Papadakos and Yannis Tzitzikas. Hippalus: Preference-enriched faceted exploration. In *EDBT/ICDT Workshops*, 2014.