半構造データに対するファセット検索に関する研究

A Study on Faceted Search for Semi-structured Data

駒水 孝裕 ♥

Takahiro KOMAMIZU

XML をはじめとする半構造データが普及しており，半構造データの利活用が重要な課題となってきている．半構造データのもつ柔軟性からしばしば複雑な構造が記述される．このような半構造データを利活用する上で，必要な情報を探しだす検索技術は重要な基礎技術の一つである．本研究では，探索的検索手法であるファセット検索に注目し，半構造データに対して探索的な検索を可能にする手法を提案する．まず，(1) 代表的な半構造データである XML データに対して，ファセット検索を行うためのフレームワークを提案する．また，(2) XML データは木構造だけでなくグラフ構造も記述可能であり，グラフ構造を有する半構造データに対してファセット検索を可能にする．(3) XML データの XML 部分木を検索対象として抽出する際に，人手が必要となる点について自動化することによってファセット検索を簡単に構築する手法を提案する．加えて，(4) XML データ内のテキストデータをファセット検索の文脈で活用する方法を提案する．

Semi-structured data such as XML data have been used in various situations in order to reuse information not only in the services but also external applications. Utilizing semi-structured data is an important challenge. This research especially focuses on the exploration over semi-structured data. Faceted search is one of the promising exploratory search methods, therefore, this research applies faceted search to semi-structured data. In order to construct a faceted search system, this research works on four directions: (1) a framework to construct a faceted search system over XML data which is a tree-structured semi-structured data, (2) the extended framework for graph-structured semi-structured data, (3) an automation scheme for extracting necessary information, and (4) utilization of textual contents in semi-structured data.

1 はじめに

半構造データは構造が明確に定義されていないデータを指し，データの公開するときにしばしば利用される．半構造データのフォーマットの例としては，XML (Extensible Markup Language) [3] や JSON (JavaScript Object Notation) [5] がある．これらのフォーマットを介して，様々なデータが公開されている．例えば，論文目録データの DBLP[1] や Twitter[2] の API で取得できるデータなどがある．半構造データに対する検索は半構造データを利活用する際の重要な基本技術である．

半構造データに対する検索において，ユーザの検索意図が明確でない場合は検索結果が多く，ユーザが満足する結果を得るまでに対話的に検索を行う必要がある．一般に検索する際のユーザの検索意図は，明瞭な検索意図と不明瞭な検索意図に大別できる．前者の場合にはユーザが検索意図を正確にクエリで表現できれば容易に必要な情報を得ることができる．しかしながら，後者の場合には必要な情報を得るために，ユーザは探索的行動を行う必要がある．探索的行動とは，(1) クエリを作成する，(2) 検索結果を吟味する，を目的の情報を得るまで繰返すことである．

ファセット検索 [10] は探索的検索手法の一つで，ユーザの探索的行動をサポートする方法である．ファセット検索において，クエリは属性名と属性値のペアの集合で表される．属性を特にファセットと呼ぶ．ファセット検索において，検索対象データにおけるファセットの値の分布を提示することで絞り込みを行うためのヒントをユーザに表示する．ユーザは自分の興味あうファセットを選択することで絞り込みを行う．また，ユーザは選択済みのファセットを排除することで検索条件を緩めることができる．ファセット検索は，電子商取引や文献目録の分野を中心に広く用いられている．例えば，Amazon [1] や eBay [2]，DBLP [6]，IEEE Xplore [4] などがファセット検索を用いている．

本研究では，半構造データに対する検索にファセット検索を導入することで，検索意図が曖昧なユーザに対して有効な検索システムを提供する．半構造データに対してファセット検索を適用するために，次の課題を解決する必要がある．ひとつは，(1) 検索対象となるオブジェクトの設定，次に，(2) オブジェクトに対するファセットの設定，最後に (3) 検索メカニズムである．本研究の貢献は以下のようになる．

- 半構造データに対するファセット検索を行うためのフレームワークを提案する．まず，代表的な半構造データである XML データに対するフレームワークを実現する手法を提案する．次に，より複雑なグラフ構造を有する半構造データに対するフレームワークを実現する手法を提案する．
- ファセット検索インターフェースを構築する際の労力を低減するために，必要な情報であるオブジェクトとファセットの自動抽出手法を提案する．
- 半構造データ中にしばしば含まれる長いテキストを検索に応

♥ 正会員 筑波大学システム情報工学研究科 taka-coma@acm.org

[1] http://dblp.uni-trier.de/
[2] https://twitter.com/

用し，効果的なファセット値を抽出する手法を提案する．
- 提案手法の被験者実験における検索タスクの決め方を提案する．探索的検索におけるタスクの探索性を表現する指標を設け，この指標に基づくタスク生成方法を提案する．

本稿の以降の構成は以下のようになる．2 節で基本事項としてファセット検索について説明する．3 節で本研究の内容について説明する．4 節で本研究における実験を紹介する．5 節で本稿のまとめと今後の課題について述べる．

2 ファセット検索

ファセット検索 [10] は属性付きオブジェクトの集合に対して，属性とその値の分布を提示することで検索結果の概要をユーザに示し，ユーザがさらに絞り込みを行うためのヒントとすることを可能にした探索的検索手法である．以下のレコード型データを用いてファセット検索について簡単に説明する．

表 1 車データベース．

車 ID	色	種類	値段
1	黒	セダン	150 万
2	赤	セダン	200 万
3	黒	SUV	300 万
4	白	SUV	200 万

表 1 のデータは，色，種類，値段，の三種類の属性をもつ．（車 ID はシステム管理用の識別子である．）ファセット検索システムはユーザに各属性（ファセット）の値の分布を提示する（表 2）．

表 2 ファセット．

(a) 色

値	頻度
黒	2
赤	1
白	1

(b) 種類

値	頻度
セダン	2
SUV	2

(c) 値段

値	頻度
200 万	2
150 万	1
300 万	1

ユーザはファセットの内容から興味についてのアイデアを得られる．ユーザは得られたアイデアにもとづいて，絞り込み条件を指定する．ここでは，“色” ファセットから “黒” を選んだとする．ファセット検索システムは車 ID 1 と 3 のレコードを検索結果としてユーザに返し，同時に，対応するファセットを提示する．これらの動作をユーザが検索結果に満足するまで繰り返す．

3 半構造データに対するファセット検索手法

半構造データに対してファセット検索を適用する際に以下に示す課題がある．

1. 半構造データにはオブジェクト及びファセットが明示されていない．ファセット検索を適用するためには，オブジェクト及びファセットを予め定義・抽出する必要がある．

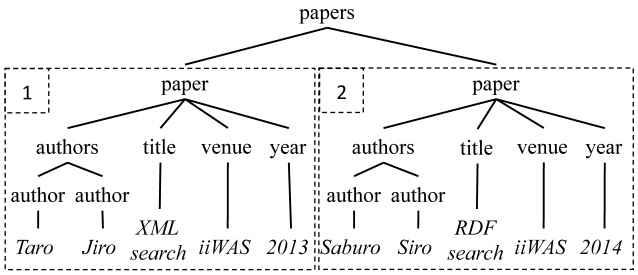

図 1 XML データ例：文献目録．

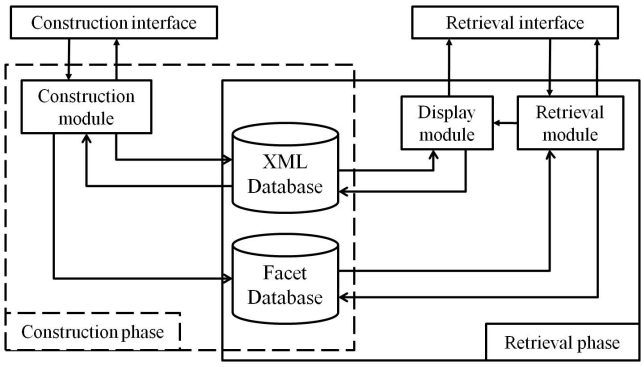

図 2 フレームワーク概要．

2. 半構造データは複雑な構造（木構造やグラフ構造）をしている．複雑な構造に対してオブジェクトやファセットの定義を手動で与えることはデータに対する高度な知識を要求する．そのため，それらの定義を与えることは容易ではない．
3. 半構造データはしばしば長いテキストを含んでいることがある．ファセット検索において長いテキストはファセット値としては有用ではない．しかしながら，長いテキストには有用な情報が含まれていることが多く，活用するべきである．

本研究はこれらの課題を解決する手法を提案する．以下の節にて，それぞれの課題に対する手法を紹介する．

3.1 XML データに対するファセット検索

木構造で表現される半構造データである XML データにファセット検索を適用するためのフレームワークを提案する．本研究が着眼した点は，XML データ内で頻出する部分木は何らかのデータのまとまりを表すという点である．図 1 に XML データの例を示す．このデータにおいて paper 要素以下の部分木（図中の点線矩形）が繰返し出現している．この頻出構造は文献を表している．提案するフレームワークはこのアイデアに基づいてオブジェクトを自動抽出する．

提案フレームワークは，構築フェイズ (Construction phase) と検索フェイズ (Retrieval phase) からなる．構築フェイズでは，検索対象となる XML データからオブジェクトおよびファセットの候補を抽出し，システム管理者に提示する．システム管理者は提示されたオブジェクトおよびファセットの候補から実際の検索

に用いるものを選別する．これによって，不要なオブジェクトやファセットを排除する．検索フェイズでは，選別されたオブジェクトとファセットに対してファセット検索を行うためのインターフェースを提供する．検索フェイズでは，ユーザの入力に応じて対応する XQuery を生成する仕組みを提供し検索結果および対応するファセットを取得し，ユーザに提示する．

3.2 グラフ構造データに対するファセット検索

木構造を有する XML データに対するファセット検索からの拡張として，グラフデータに対するファセット検索を可能にするフレームワークを提案する．実世界データを有するグラフデータは，個々のデータを似たような構造で持っていることが期待される．このアイデアに基づき，頻出部分グラフマイニング [7, 11] を用いて頻出する部分グラフをオブジェクトとして抽出，ファセット検索を実現する手法を提案する．フレームワークの基本骨格は，図 2 と同様であり，オブジェクト及びファセットの抽出に頻出部分グラフマイニングの技術を用いる．

3.3 オブジェクト及びファセット自動抽出

提案フレームワークにおいて，システム管理者はオブジェクト及びファセットの候補から適切なものを選出しなければならない．この主目的は，不要なオブジェクト及びファセットをユーザインターフェースに提示しないようにするためである．しかしながら，特にデータが巨大で構造が複雑であるとき，この作業はシステム管理者にとって大きな負担となる．

本研究では，システム管理者の負担を減らすべく，オブジェクトおよびファセットを自動で選出するヒューリスティックアプローチを提案する．まず，オブジェクトについては，不必要に頻出になってオブジェクトとして検出されるエラーに着目し，安定して頻出する構造をオブジェクトとして抽出する手法を提案する．次に，ファセットについては，ID のような振舞いをするファセット値を有するファセットが有用でないことが多い事実に着目し，ファセット値の頻度分布に基づく自動選出手法を提案する．最後に，オブジェクトおよびファセットについて，選出される要素が人にとって有意である点に着目し，辞書に含まれる単語を含む要素をオブジェクト及びファセットとして抽出する．

3.4 テキスト値からのファセット値の抽出

提案フレームワークにおいて，長いテキストをファセット値としてもつファセットは，システム管理者によって，あるいは，自動的に排除されてきた．これは，長いテキストがユニークな値であり，ファセット検索にとって有用でないからである．例としては，文献データにおけるタイトルがある．しかしながら，長いテキストにも有用な情報が含まれている．文献のタイトルには文献の内容を示す内容が書いてあるはずである．本研究ではこのような内容を表すファセット値を抽出し，より有用なファセット検索システムを提供する．

本研究では，テキストから内容を表し絞り込みに向いている単語を抽出する．このよう条件を満たす単語を抽出するために，概念階層を抽出する手法である Subsumption [9] を用いる．

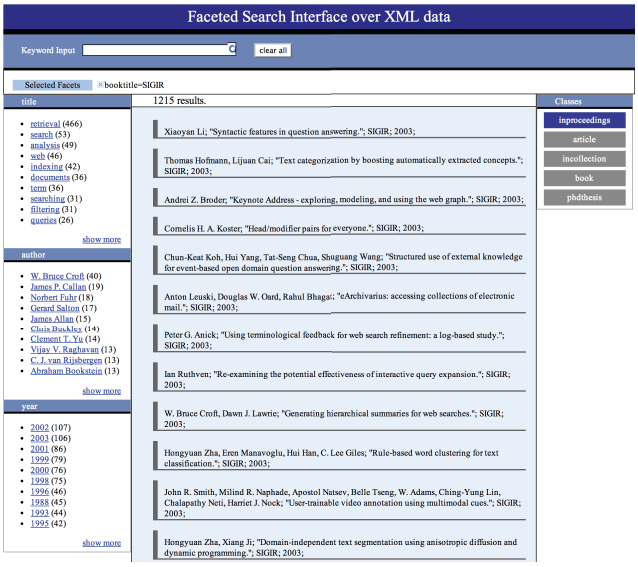

図 3 検索インターフェース．

Subsumption は単語の共起関係から単語の概念的上下関係を導出する手法である．結果として得られた概念階層の上層部をファセット値として利用することで，オブジェクト集合を徐々に絞り込むことを可能にする．

図 3 に本研究で提案するフレームワークのユーザインターフェースを示す．ユーザインターフェースは四つのコンポーネントからなる．一つはキーワード検索を行うためのコンポーネント（図中上部），もう一つはファセットと値を表示するコンポーネント（図中左部），もう一つはオブジェクトの種類（クラス）を表示するコンポーネント（図中右部），最後は検索結果のオブジェクトを表示するコンポーネント（図中中央）である．この図は，DBLP の XML データに対して本フレームワークを適用した際の検索インターフェースである．テキストからのファセット値抽出を行ったファセットは title である．故に，図中の title のファセット値は単語からなっている．

4 実験

本研究の有用性を示すために，以下の実験を行った．

- XML データに対する典型的な問合せ言語 XQuery を用いた場合とファセット検索を用いた場合の比較．
- グラフデータに対する適用可能性を示す事例検証．
- オブジェクトとファセットの自動抽出の精度実験．
- テキストからファセット値を抽出した場合の有用性を検証するための被験者実験．

本稿ではスペースの都合上，最後の項目についてのみ説明する．

本研究で行った被験者実験は，ユーザに検索タスクを与え，ユーザがタスクを達成するまでにかかった時間を計測するというものである．そのためにまず，タスクを設計する必要がある．タ

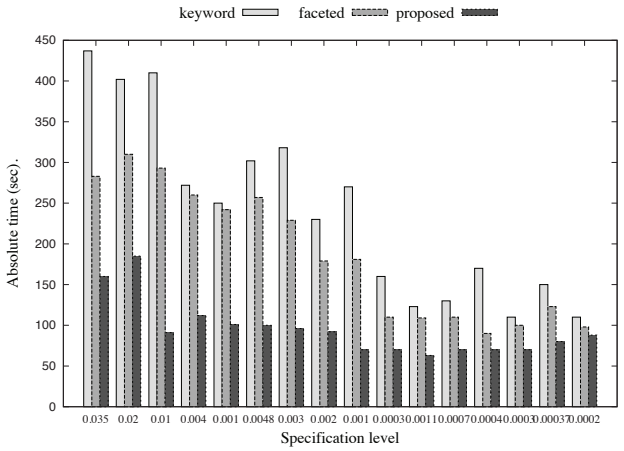

図 4　被験者実験：タスク達成時間の分布

スクの設計にはテンプレートに基づく設計方法 [8] を元にテンプレートの空欄を埋める語句を選出する方法を提案する．探索的検索の評価を行う上で，タスクがどの程度 “探索的” であるかを示す指標が必要である．これをタスクに含まれる語句の選択率ととらえた指標 specification level を提案し，specification level に基づいてタスクの空欄を埋める方法を提案する．

実験で用いたデータは DBLP の XML データで，title をテキストからファセット値抽出するファセットとした．フレームワークを用いることで，文献を表す部分 XML 木 (proceedings や book など) をオブジェクトとして抽出した．

図 4 に探索度合いを変えた時のユーザのタスク達成時間の分布を示した．横軸は specification level で左に行くほど探索度合いが高く，右に行くほど探索度合いが低くなる．比較対象は，テキストからファセット値を抽出せずにファセット検索を行った場合 (faceted, 白色) とキーワード検索 (keyword, 灰色) である．図から提案手法が最も効率的に検索出来ていることが読み取れる．特に探索度合いが高いほど，効果的であるといえる．

本実験に用いた specification level の有効性を確かめるため，specification level と実験データの相関係数を観測した（表 3）．

表 3　各種指標と specification level の相関係数.

measurement	keyword	faceted	proposed
time	0.771346	0.655485	0.841967
# operations	0.236971	-0.411838	-0.122954
# keywords	0.236971	-0.39138	-0.063875
# facets	—	-0.299988	-0.122954

この表から specification level と時間 (time) に強い正の相関があり，specification level を変化させた時に探索時間に影響が強いことが明らかになった．一方で，キーワード検索の利用数やファセットの選択数とは相関があまりないことが観測された．故に，ファセットの選択数を基準とした被験者実験では specification level は意図した探索度合いを表現しにくいことが明らかになった．しかしながら，本研究は速やかに目的（タスク）を達成できるファセット検索インターフェースを良い検索インターフェースとかんがえるため，本実験は有意義であるといえる．

5　まとめと今後の課題

本稿では，半構造データに対するファセット検索に関する研究について提案した．データ指向の半構造データがもつ特性に注目して，ファセット検索に必要であるオブジェクト及びファセットを抽出する手法を提案した．今後の課題は，(1) データが分散して存在する場合にファセット検索を行う手法の検討，(2) データ分析と組み合わせた手法，などの応用研究が考えられる．加えて，探索的検索の評価方法についても研究の余地があると考える．

[謝辞]

筑波大学の北川博之教授，天笠俊之准教授を初めとする諸先生方やご指導くださった皆様に深く感謝致します．

[文献]

[1] Amazon.com. http://www.amazon.com/.
[2] eBay. http://www.ebay.com/.
[3] Extensible Markup Language (XML). http://www.w3.org/XML/.
[4] IEEE Xplore Digital Library. http://ieeexplore.ieee.org/Xplore/home.jsp.
[5] JSON (JavaScript Object Notation). http://www.json.org/.
[6] The DBLP Computer Science Bibliography. http://www.informatik.uni-trier.de/~ley/db/.
[7] S. Ghazizadeh and S. S. Chawathe. SEuS: Structure Extraction Using Summaries. In *Discovery Science*, pages 71–85, 2002.
[8] B. Kules, R. Capra, M. Banta, and T. Sierra. What do exploratory searchers look at in a faceted search interface? In *JCDL*, pages 313–322, 2009.
[9] M. Sanderson and W. B. Croft. Deriving Concept Hierarchies from Text. In *SIGIR*, pages 206–213, 1999.
[10] D. Tunkelang. *Faceted Search*. Synthesis Lectures on Information Concepts, Retrieval, and Services. Morgan & Claypool Publishers, 2009.
[11] F. Zhu, Q. Qu, D. Lo, X. Yan, J. Han, and P. S. Yu. Mining Top-K Large Structural Patterns in a Massive Network. *PVLDB*, 4(11):807–818, 2011.

駒水 孝裕 Takahiro KOMAMIZU
2011 年筑波大学システム情報工学研究科博士前期課程修了．2015 年筑波大学システム情報工学研究科博士後期課程修了．博士（工学）．ACM 会員．日本データベース学会正会員