D-022 XML データに対するファセット検索のためのファセット抽出の自動化
Automating Facet Extraction for Faceted Search over XML Data

駒水 孝裕 * 天笠 俊之 † 北川 博之 †
* 筑波大学大学院システム情報工学研究科 † 筑波大学システム情報系
*taka-coma@acm.org †{amagasa,kitagawa}@cs.tsukuba.ac.jp

I. はじめに

ユーザは XML データに対する検索を行う際に，検索対象の XML データの構造及び内容についてある程度の知識を持っていることが要求される．代表的な検索方法に，(1) XPath を用いた検索と (2) キーワード検索がある．(1) は XML データ内の検索対象要素までのパス及び条件を記述し検索する方法で，(2) はいくつかのキーワードを入力しそれらを含む部分木を検索する方法である．より正確な結果を得るためには，ユーザはより細やかなクエリを上記記法にしたがって記述しなければならない．しかしながら，XML データについての知識が専門的でないユーザにとって精確なクエリを記述することは容易ではない．

ファセット検索 [1] はユーザの探索行動をサポートする探索的検索手法 [2] の一つで，ユーザに検索結果の概要をファセットの形で示し，ユーザは提示されたファセットを選択することで検索結果の絞込みを行える．ファセット (*facet*) とは，検索対象データの側面を表現するもので，オブジェクトに対する属性やカテゴリのようなものである．人物情報を例に取ると，ある人物は次のようなファセットと値を持つ．

(名前, 筑波太郎), (年齢, *27*), (職業, カメラマン), (性別, 男)

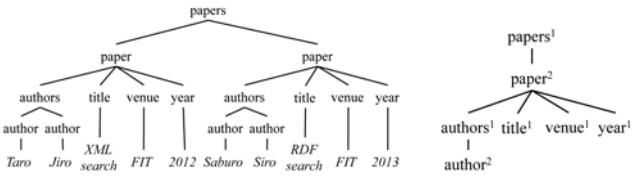
この例において，ファセットは { 名前, 年齢, 職業, 性別 } であり，それぞれに組付けされている文字列はそれぞれのファセットに対する値である．ユーザが人物に関する情報を検索したい場合には，いずれかのファセットを選択し，提示された候補となる値を選択することで検索結果を絞り込む．ファセット検索では，この選択行動をユーザが望む結果を得るまで繰り返す．先行研究 [3] にて，我々は XML データに対するファセット検索を可能にするフレームワークを提案した．

本稿では，先行研究 [3] で提案した手法における XML データから探索対象オブジェクトおよびファセットの抽出プロセスの自動化に取り組む．XML データに対するファセット検索を行う上で，まず，どの XML 部分木を検索対象オブジェクトかを決める必要がある．先行研究では，ある XML 要素と同名の要素がその親要素に複数回出現する場合にそれをオブジェクトの候補とし，最終的には人手で検索対象オブジェクトを決めていた．また，ファセットは検索対象オブジェクトとなる XML 要素以下の XML 要素でテキストノードを直に持つものをファセットの候補とし，同様に人手でファセットを決めていた．これらに対し，本研究ではこれらの検索対象オブジェクト及びファセットの選定にかかる検索インターフェース設計者の負荷を減らすべく，これらのプロセスの自動化を行う．

II. 問題定義

A. 基本概念

a) 構造要約*:* XML データに対する構造要約とは，XML データの取りうる構造を記述したものであり，XML データを走査することで得られる．構造要約を見ることで XML データ中に存在する XML 要素の取りうる親子関係や親要素に対する子要素の出現頻度などを知ることができる．構造要約は XML 同様に木構造で表現される．Fig. 1(a) の例示 XML データに対する構造要約を Fig. 1(b) に示した．本研究で用いる構造要約では，それぞれのノードに XML データ内での親要素以下での平均頻度をラベルとして付加している．



(a) XML データ. (b) 構造要約.

Fig. 1. 例示 XML データと構造要約.

b) クラスとファセット*:* 構造要約内のあるノードで XML データ中のオブジェクトに対応するものをクラスと定義する．構造要約中のクラスノードの子孫ノードの内，選ばれたノードをファセット定義する．

B. 本研究の目的

与えられた XML データから抽出した構造要約からクラスとファセットを自動的に抽出することが本研究の目的である．

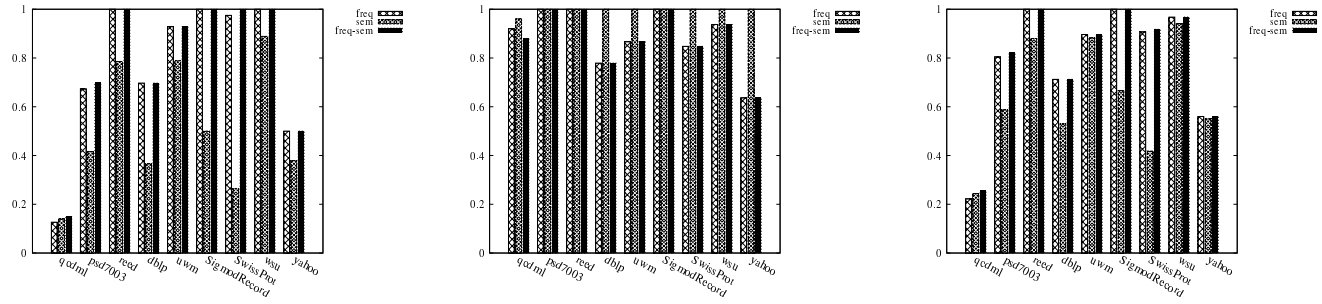
III. フレームワーク

与えられた XML データから検索対象オブジェクトとファセットを抽出するために必要なプロセスを以下のようなフレームワークで行う．(1) XML データから構造要約を抽出する．(2) 構造要約からクラス候補及びファセット候補を抽出する．(3) クラス候補及びファセット候補から適切なものを抽出する．(4) XML データから抽出されたクラスに該当する XML 部分木をオブジェクトとして抽出する．(5) 抽出されたファセットに該当する XML 要素の値をファセットの値として抽出する．本稿は，上記プロセスのうち (2) および (3) の自動化に取り組むものである．

本研究のアプローチは経験則に基づき，クラス及びファセットを自動的に抽出する手法である．基本的なアイデアは以下の三点である．(1) より頻繁に出現する XML 要素は検索対象オブジェクトとなりやすい．(2) 要素のテキスト値が検索対象オブジェクトの対して一意にきまるような XML 要素はファセットとなりにくい．(3) 人が意味を解釈できないような機械的な XML 要素は検索対象オブジェクト及びファセットとなりにくい．(1), (2) を基にした手法を頻度に基づくアプローチ，(3) に基づく手法を意味に基づくアプローチと呼称する．

A. クラスの抽出

XML データ中で単一の親 XML 要素の下で複数回出現するような XML 要素はある情報の単位を表していると考えられる．単純に複数回出現する XML 要素をクラスとみなす方法 [3] は不要な XML 要素までクラスとみなす可能性がある．例えば，下付き文字を表す `sub` 要素は文書を表現する `description` 要素の下で複数回出現することがある．しかし，`sub` 要素は文書の構成に係る要素であり，それ自身が情報のまとまりを表すことは稀である．このような文書の構成に係る要素は，文書によって使われないこ



(a) Precision (b) Recall (c) F-Score

Fig. 2. ファセット抽出の精度.

ともあり，同名の親要素の下での平均出現回数が低い．このような観測から，頻度に基づくクラスの抽出手法では，以下のようにして平均出現頻度が閾値を超えるような XML 要素をクラス集合 C^f として抽出する．

$$C^f = \{v \,|\, v \in V \wedge avg\_freq(v) > \theta_c\} \quad (1)$$

ただし，V は XML データ中のユニークな XML 要素の集合を表し，$avg\_freq(v)$ は要素 v の親要素以下での平均出現頻度を返す関数であり，θ_c は出現頻度の閾値である．

一方で，XML 要素名が意味的に解釈可能でない XML 要素は検索対象オブジェクトとなりにくいと考えられる．そこで，既存の知識（WordNet や Wikipedia）を用いて，XML 要素名が意味的に解釈可能であるかを判定し，可能な場合にクラスとして抽出する．その際に，XML 要素名に含まれる単語のゆらぎに対応するために，意味的に類似した単語が知識に存在する場合にクラスとして抽出するように拡張を行う．この手法を意味に基づくクラス抽出手法と呼び，以下のように定式化する．

$$C^s = \{v \,|\, v \in V \wedge \exists i \in I(sem\_sim(v,i) > \theta_{cs})\} \quad (2)$$

ただし，I は既存の知識の集合，$sem\_sim(v,i)$ は v と i の意味的類似度を計算する関数 [4]，θ_{cs} は意味的類似度の閾値である．

頻度に基づくアプローチと意味に基づくアプローチは互いに独立な手法であるため，これらを合わせることは容易であり，合わせた手法をハイブリッド手法と呼ぶ．ハイブリッド手法は以下のようにクラス集合 C を計算する．

$$C = C^f \cap C^s \quad (3)$$

B. ファセットの抽出

検索対象オブジェクトを検索するためのファセットはオブジェクトの識別力の高くないものが良いと考えられる．すべての子孫要素をファセット候補とする手法 [3] はテキスト値が検索対象オブジェクトに対して一意に決まる XML 要素もファセット候補として抽出していた．このような XML 要素はファセットとしては不的確である．このような XML 要素をファセットとして抽出することを避けるため，XML 要素のテキスト値の平均出現頻度が一定以上のものをファセットとして抽出する．

$$F_c^f = \{v \,|\, v \in desc(c) \wedge avg\_val\_freq(v) > \theta_a\} \quad (4)$$

ただし，$desc(c)$ はクラス c の構造要約上の子孫ノードの集合を返す関数で，$avg\_val\_freq(v)$ は XML データ中において XML 要素 v のテキスト値の平均出現頻度を計算する関数で，θ_a は出現頻度の閾値である．

一方で，XML 要素名が意味的に解釈可能でない XML 要素は検索対象オブジェクトとなりにくいと考えられる．そこで，クラス抽出と同様に既存の知識を用いて解釈可能な要素名をもつ XML 要素をファセットとして抽出する．

$$F_c^s = \{a \,|\, a \in desc(c) \wedge \exists i \in I(sem\_sim(a,i) > \theta_{fs})\} \quad (5)$$

ただし，θ_{fs} は意味的類似度の閾値である．

頻度に基づくアプローチと意味に基づくアプローチは互いに独立な手法であるため，これらを合わせることは容易であり，合わせた手法をハイブリッド手法と呼ぶ．ハイブリッド手法は以下のようにファセット集合 F_c^h を計算する．

$$F_c^h = F_c^f \cap F_c^s \quad (6)$$

IV. 評価実験

提案手法の有効性を示すために，我々は正解データを作成し，クラス及びファセットの検出精度を計測する実験を行った．実験に使用したデータは，UW XML Repositories のデータ [1] と QCDml[2] を用いた．意味に基づくアプローチの知識ベースは Wikipedia のエントリを用いた．評価指標としては，適合率 (Precision)，再現率 (Recall) と F 値 (F-score) を用いた．スペースの都合で本稿では，ファセットの検出精度についてのみ言及する．Fig. 2 にファセットの検出精度の結果を表示した．`freq` は頻度に基づくアプローチを，`sem` は意味に基づくアプローチを，`freq-sem` はハイブリッドアプローチを表す．

実験結果から，頻度に基づくアプローチが十分な精度を示し，意味に基づくアプローチがそれを補強することが見て取れる．意味に基づくアプローチ単体では，多量の XML 要素をファセットとして抽出してしまうため，過剰に抽出してしまう傾向にある．頻度に基づくアプローチと組み合わせることで，頻度だけでは判断できない XML 要素をファセットから除外できている．

V. まとめと今後の課題

本稿では，XML データに対するファセット検索を行うための検索対象オブジェクト及びファセット抽出の自動化に取組んだ．本稿では，頻度に基づく手法と単語の意味に基づく手法を提案した．また，実験により，閾値を変化させた際の影響について評価を行った．今後の課題としては，XML 要素の中身の分析を取入れて判別するなどのより高度な情報を取込むことに精度改善があげられる．

謝辞

本研究開発の一部は文部科学省委託事業「実社会ビッグデータ利活用のためのデータ統合・解析技術の研究開発」による．

参考文献

[1] D. Tunkelang, *Faceted Search*, ser. Synthesis Lectures on Information Concepts, Retrieval, and Services. Morgan & Claypool Publishers, 2009.
[2] R. W. White, B. Kules, S. M. Drucker, and M. Schraefel, “Supporting Exploratory Search,” *Commun. ACM*, vol. 49, no. 4, 2006.
[3] T. Komamizu, T. Amagasa, and H. Kitagawa, “A Framework of Faceted Navigation for XML Data,” in *Proc. iiWAS*. ACM, 2011, pp. 28–35.
[4] S. Harispe, S. Ranwez, S. Janaqi, and J. Montmain, “Semantic Measures for the Comparison of Units of Language, Concepts or Entities from Text and Knowledge Base Analysis,” *CoRR*, vol. abs/1310.1285, 2013.

[1] http://www.cs.washington.edu/research/xmldatasets/
[2] http://www.jldg.org/facetnavi/