メタサーチ環境におけるファセット検索の実現

諫本　有加　（指導教員：渡辺知恵美）

1　はじめに

今日 e-science の分野おいて，科学データは大規模な研究組織だけでなく，各研究者が個々に研究経過や成果として所有・公開している．そのため，世界中に散在される情報を集めて検索できるように科学データのメタサーチエンジンの需要が高まっている．各メタサーチエンジンは独自の問合せパラメタを持っているため，これらのサーバのデータに対してまとめて検索を行う場合，検索インタフェースとしてキーワード検索が考えられる．しかしながらユーザが考えたキーワードで検索しても求めるデータを取得できないなど，所望のデータに辿り着くための適切なキーワードを見つけることは難しい．本研究では，データの種類やデータがもつ情報を知らなくても提示された要素を選択するだけで求めるデータに辿り着く検索インタフェースであるファセット検索に着目し，それをメタサーチ環境で実現するための手法を提案する．本研究では検索対象とするサーバの数に合わせ 2 つの手法を提案した．1 つ目は REST ベースの Web サービスを通して問合せを実現する手法である．これは対象サーバがいくつかに特定されている場合に有効である．2 つ目は P2P ネットワークを用いて問合せを行う手法であり，これは対象サーバが特定されていない場合に利用できる．紙面の関係上，本稿では後者の P2P 上でファセット検索を実現するための技術提案および検証結果についてのみ述べる．

2　ファセット検索

ファセット検索とはデータの検索条件として属性やメタデータ項目を予めリスト表示しておき，それを選択することで目的のデータに辿り着けるインタフェースである．ファセット検索ではデータを絞り込むための条件をファセットと呼び，それぞれのファセットを分類したクラスの名前をファセット名，その中でリストアップされた条件をファセット値と呼ぶ．

ファセット検索の代表例である Flamenco Search [1] とそこで使われているデータを分かりやすく図式化したものを図 1 に示す．

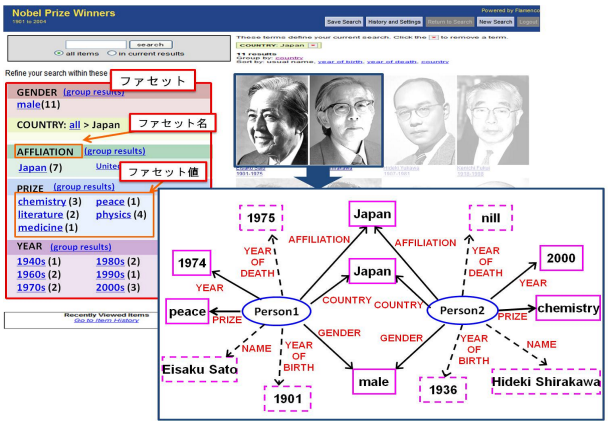

図 1: Flamenco Search のデータ

図 1 は「country:Japan」で検索した結果である．この条件に該当するオブジェクト person1 と person2 は「gender:male」などの他の属性をもっており，これらはファセットとしてリストアップされる．ユーザは提示されたファセットの中から次の条件を選択することで，繰り返し検索を行うことができる．

3　P2P におけるファセット情報の配置

先行研究 [2] では P2P 上でファセット検索を実現するためのデータの配置方法を分散ハッシュテーブル (DHT) を想定して提案した．

ファセット検索では条件に該当するオブジェクトがもつその他のファセットを取得する必要がある．そこで我々は，検索条件となるファセット名とファセット値の組を基にハッシュ値を求め，その値に対応するノードに検索条件にあうオブジェクトが他にもつファセット名とファセット値を格納する．例えば,「coutry:Japan」に該当する person1 と person2 がもつ他のファセットを格納するためには「country:Japan」のハッシュ値を求め，該当オブジェクト，ファセット名，ファセット値のリストを格納する（図 2）．

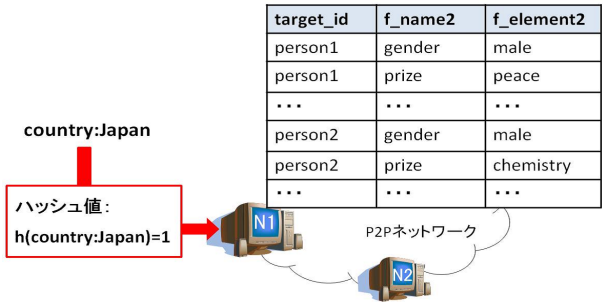

図 2: ファセット情報の配置

4　P2P におけるファセット問合せ処理

4.1　問合せ処理の流れ

P2P におけるファセット検索の処理の流れを以下に示す．

Step1. ユーザが条件を指定すると，その条件を key にして求められたハッシュ値と一致するノードに問合せ，該当タプルを絞り込む

Step2. 集約結果としてファセット名とファセット値，さらに該当件数を取得する

Step3. 取得した結果を表示し，その中からユーザが次の絞込み条件を選択する

4.2　処理ノードの検討

Step2 において以下の 2 通りの手法が考えられる．

手法 (a)： ノード側で集約処理を行い，クライアントは結果のみ取得する

手法 (b)： ノードから該当タプルをそのまま取得し，クライアント側で集約処理を行う

文献 [3] において，それぞれの手法での処理時間とユーザが選択しやすい条件の有無を検証した．処理時間の結果を図 3 に示す．

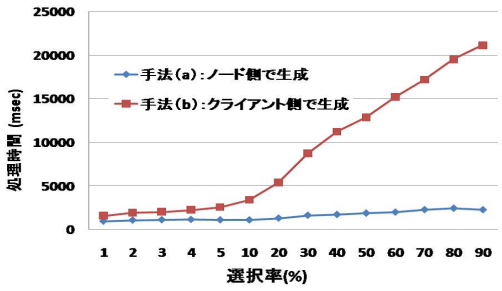

図 3: 処理時間の比較

図 3 より，ノード側で集約処理を行う手法 (a) の方が処理時間が速い．これは，手法 (a) ではクライアント側は結果のみを取得することから，ネットワークトラフィックが抑えられるためである．また [3] より，2 つの手法の処理時間はクライアントのネットワーク状況にかかわらず，ノード側で処理を行う手法 (a) が適切であるとわかった．また選択する条件の検証において，ユーザが選択する条件には偏りがあり，さらにその条件を複数回利用するという結果を得た．

4.3 問合せ対象ノードの検討

文献 [3] で行った予備実験より，ユーザに選択されやすい条件があり，さらにその条件は繰り返し利用されることがわかった．当初，同じノードに問合せを行うとキャッシュにデータが残っており処理が速くなると考え，1 つ目の条件を key に固定することを想定していた．そのため，図 4 のように複数のユーザが最初に条件 A を指定して検索を繰り返すと条件 A に該当するファセットをもつノードに問合せが集中する．

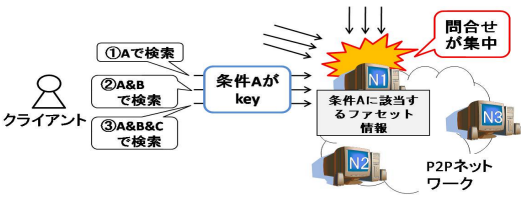

図 4: 問合せが集中する状況

そこで我々は，条件が複数になった場合に問合せごとに key を変えることで，特定のノードに処理が集中しないように以下の 2 つの key 選択の手法を提案する．

手法 (2)： 問合せを行う key をランダムに抽出する

手法 (3)： ノードへのアクセスがそのノードの許容量に達した場合，次の条件を key にする

手法 (2) ではランダムに key を抽出するため，問合せが多くのノードに分散され，1 ノード辺りの問合せ数が少なくなると考えられる．手法 (3) では，1 つ目の条件として頻繁に利用される条件を key にもつノードへの問合せが多くなっても許容量を超えることがないため，問合せ集中による処理能力の低下などを抑えることができる．ただし，問合せが複数のノードに一定に分散されず，偏りが出てしまう．

4.4 検証

3 つの手法に対して問合せ集中と処理時間の検証を行った．20 個のノードに対して 300 のクライアントが問合せを行うとし，その際、複数のクライアントから 1 つ目に選択される条件があること，その条件が複数回繰り返し利用されることを想定した．それぞれのノードのアクセスの許容量は 500m 秒毎に 20〜50 とした．それぞれの比較結果を以下に示す．

	許容量を超えた割合
当初の方法	40.9%
手法(2)	10.1%
手法(3)	0%

図 5: 問合せ集中の結果

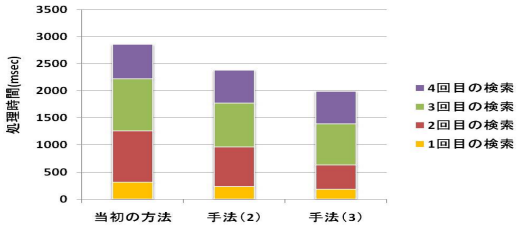

図 6: 処理時間の比較

図 5 より，当初の方法と手法 (2) で問合せの集中が見られた．当初の方法では選択されやすい条件を key にもつノードに問合せが集中し，許容量を超える．手法 (2) では key の選択をランダムにしても，非常に多く出現する条件は選択される可能性が高いことから許容量を超えることがある．図 6 より，当初の方法と手法 (2) で処理時間が大きい．これは，2 つの手法は問合せの集中が起こることで処理が遅くなるためである．したがって 2 つの検証結果より，手法 (3) がノードの負荷分散に適切であるといえる．

5 まとめと今後の課題

本稿ではメタサーチ環境におけるファセット検索の実現する手法を述べた．その際，対象サーバの数に対して 2 つの手法を提案した．対象サーバが特定されている場合のメタサーチとして Web サービスを用いてファセット検索を実現する手法を提案し，実装を行った．対象サーバが大規模である場合は P2P 環境を利用し，効率的なファセット検索を実現するための技術提案および検証を行い，その結果について述べた．今後は検証結果を基に実装を行っていきたい．

参考文献

[1] Flamenco Search ”http://flamenco .berkeley.edu/index.html”

[2] 齋藤真衣，渡辺知恵美：”P2P 環境における RDF データを対象にした Faceted Search の実現”，第 2 回データ工学と情報マネジメントに関するフォーラム DEIM(2010), C9-4(2010)

[3] 諫本有加，渡辺知恵美：”P2P 環境における Faceted Search の実現に向けて”，情報処理学会報告，2010-DBS-154