4J－5

分散メディエータ環境における索引情報のキャッシング手法の提案

片山薫　石川佳治　植村俊亮
奈良先端科学技術大学院大学 情報科学研究科

1. はじめに

インターネットの普及に伴い，大量の情報がネットワーク上に流通するようになり，その中から必要な情報を取り出すための仕組みが求められている．ネットワーク上に分散する，様々なインタフェース，セマンティクスを持つ情報源を統合し，統一したインタフェースを提供することで，利用者による情報源へのアクセスを仲介するソフトウェアをメディエータ (mediator) と呼ぶ．

本研究室では，考古学分野を対象として，各研究機関で作成されたデータなどの流通や統合に関する研究を行なっている[4]．本稿では，多くのサイトで提供される報告書などの文書に対するアクセスをサポートするためメディエータを用いることを想定し，その検索処理を効率化するための手法として，1) で提案されたシグネチャキャッシュを利用する手法を提案する．ここでのメディエータの役割は，文書情報を提供する個々の情報源からその内容を抽出して索引情報 (シグネチャ) を作成し，検索情報の流通を仲介することである．クライアント (利用者側の計算機) は，メディエータから返される検索結果を，その索引情報と共にキャッシュし，後で関連する問合せを行なう時に再利用することで，処理を効率化することができる．

2. シグネチャキャッシュ

シグネチャキャッシュ(SignatureCache) は，情報資源を検索する各クライアントで管理される一種のシグネチャファイル (signature file) である[1]．以下では，シグネチャキャッシュの概念と，その基となったシグネチャファイルについて述べる．

2.1 シグネチャファイル

シグネチャファイルは，主に情報検索の分野で用いられてきた索引手法である[5]．シグネチャ(signature) は，個々の文献の内容をコンパクトに表現した固定長 (Fビットとする) のビット列で，一般的にはスーパーインポーズドコーディング (superimposed cording) と呼ばれる手法を用いて生成される．スーパーインポーズドコーディングではまず，文献中にあらわれるそれぞれの単語を，ハッシュ関数によって，Fビットの単語シグネチャ(word signature) に変換する．単語シグネチャでは，Fビットのうちパラメータ m で決められた数だけ'1' が立てられる．次に，文献中の各単語について生成された単語シグネチャの各ビットごとに論理和をとることによって，文献シグネチャ(document signature) を生成する (図 1)．各文献の，文献シグネチャと文献番号の組を格納したものがシグネチャファイルである．

単語	単語シグネチャ
database	1000100000010100
distributed	0010110010000000
文献シグネチャ	1010110010010100

図 1　シグネチャの生成 (F=16,m=4)

シグネチャファイルを用いた検索では，まず問合せとして与えられた単語の集合から文献シグネチャの作成と同様の方法で問合せシグネチャ(query signature) を作成する．次に，問合せシグネチャとシグネチャファイル中の文献シグネチャとのパターンマッチが行なわれる．文献シグネチャは，問合せシグネチャにおいて'1' を持つビットの位置全てについて'1' の値をとる時，問合せを満たす候補となり，ドロップ (drop) と呼ばれる．しかし，ハッシュ関数の衝突やーパーインポーズドコーディングの影響で文献シグネチャが誤ってパターンマッチしてしまうことがある．このように，ドロップのうち実際には問合せ条件を満たさないものをフォルスドロップ (false drop) と呼び，満たすものをアクチュアルドロップ (actual drop) と呼ぶ．

2.2 シグネチャキャッシュの概念

シグネチャキャッシュは，クライアントにおいて発行された問合せに対し，その問合せ結果 (文献 ID) だけでなく，各文献に対応したシグネチャをキャッシュする索引情報のキャッシング手法である．索引情報に加え，問合せ結果が実際にはアクチュアルドロップであったかどうかといった情報も管理される．

例をもとにシグネチャキャッシュを用いた問合せ処理を説明する．クライアントで以下の問合せ Q1「database と distributed が含まれる文献を検索せよ」が発行されたとする．現在，シグネチャキャッシュには何も入っていないとする．

```
Q1: FIND database distributed
```

Q1 に対する検索結果がシグネチャキャッシュに保管された後，以下の問合せ Q2「database と distributed, retrieval が含まれる文献を検索せよ」が発行されたとする．

```
Q2: FIND database distributed retirieval
```

Q2 は Q1 に包摂されるため，検索結果は既にシグネチャキャッシュに保管されていることになる．よってクライアントはメディエータへ問合せを行なうことなく，シグネチャキャッシュを検索することによって問合せに答えることができる．

An Index Caching Method for Distributed Mediated Environment
Kaoru KATAYAMA, Yoshiharu ISHIKAWA and Shunsuke UEMURA
Graduate School of Information Science,
Nara Institute of Science and Technology (NAIST)

3. 文書メディエータのアーキテクチャ

本研究で想定する文書メディエータのアーキテクチャを図2に示す.

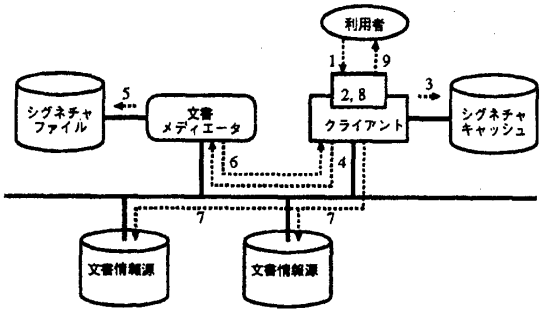

図2 文書メディエータ

文書メディエータは文書情報源に定期的にアクセスし, その最新の内容についての知識を索引 (シグネチャファイル) として保持する. クライアントからの文書の検索要求に対し, その問合せシグネチャを用いてシグネチャファイルを検索することで条件を満たす文書IDを取り出し, その索引情報と共にクライアントに返す. メディエータによる情報源へのアクセス頻度は, 各情報源の特性 (更新頻度など) やクライアントからの要求に応じて変更できる.

クライアントは利用者に対して, 文書検索のためのユーザーインタフェース (キーワードの指定) を提供する. 文書メディエータと同じハッシュ関数を保持し, キーワード列を問合せシグネチャに変換して文書メディエータへ問合せを行なう. 文書メディエータから返された検索結果は, それがアクチュアルドロップかどうかをチェックした後利用者へ返される. これらの結果はシグネチャキャッシュに保管され, 以後の問合せの最適化に用いられる.

4. 問合せ処理

ユーザーの問合せは以下のように処理される.

1. 利用者が問合せキーワードのリストを指定する.
2. クライアント側で問合せシグネチャを作成する.
3. 過去の問合せ結果を利用できないかどうかシグネチャキャッシュをチェックする.
4. シグネチャキャッシュには含まれない情報についてメディエータへ問合せを発行する.
5. メディエータで問合せを受けとり, シグネチャファイルの検索を行なう.
6. マッチした文書IDを, 文献シグネチャと合わせてクライアントへ返す.
7. クライアントは文書IDを基に文書へアクセスする.
8. アクセスした文書が問合せ条件を満たしているか (アクチュアルドロップ) かどうかをチェックし, シグネチャキャッシュにその情報を付加する.
9. 問合せ条件を満たした文献を利用者に返す.

5. キャッシュの更新処理

シグネチャキャッシュは, 問合せ時に得られたメディエータが管理するシグネチャファイルの一部を持っている. このようなアーキテクチャでは, クライアントにおけるキャッシュ内容とメディエータが管理するシグネチャファイルとの整合性の管理が重要となる. このような分散環境におけるキャッシュの更新戦略については, 以下のような方針が考えられる[3].

遅延更新 クライアントがシグネチャキャッシュを用いる時, 同時にメディエータへシグネチャキャッシュの内容が最新のものであるかどうかを問い合わせる. キャッシュ内容が古い場合は, キャッシュ内容を更新した後, 問合せを処理する.

定期的更新 定期的にクライアントがシグネチャキャッシュを更新する.

閾値による更新 メディエータは自身のシグネチャファイルの更新回数を監視し, 予め設定した閾値を越えたらシグネチャキャッシュを更新するようクライアントに指示する.

即時更新 メディエータ側においてシグネチャファイルが更新されると同時にシグネチャキャッシュを更新するようクライアントに指示する.

6. おわりに

本稿では, 分散メディエータ環境における文書情報の流通の効率化を目的とした, 索引情報のキャッシング手法について論じた. メディエータ環境におけるキャッシング手法としては, 問合せとその統計情報をキャッシュするアプローチが提案されているが[2], 本研究とは索引情報をキャッシングする点で異なる. シグネチャファイルの構成が単純であることにより, 部分的なキャッシングなどが可能であると考えられる. 今後は性能評価や更新戦略などについて研究を行なう予定である.

謝辞

本研究を進めるにあたり御指導・御討論を頂いた植村研究室の皆様に感謝いたします. また, 御助言等を頂いた筑波大学北川博之助教授に御礼申し上げます.

参考文献

1) 石川佳治, 北川博之. シグネチャキャッシュ–分散環境における情報資源アクセスの効率化手法–. 情報処理学会第52回全国大会, 1996年3月.

2) S. Adali, K.S. Candan, Y. Papakonstantinou, V.S. Subrahmanian. Query Caching and Optimization in Distributed Mediator System. *Proc. ACM SIGMOD Conf.*, pages 137-148, 1996.

3) Nick Roussopoulos, Nikos Economous, and Antony Stamenas. ADMS: A Testbed for Incremental Access Method. *IEEE Transactions On Knowledge and Data Engineering*, VOL. 5, NO. 5, October 1993.

4) 古舘丈裕, 岡安光彦, 石川佳治, 植村俊亮. 構造化文書データベースに対するラッピング手法の提案. 情報処理学会データベースシステム研究会研究報告, 1996年7月.

5) C. Faloutsos. Signature Files. in W.B. Frakes and R. Baeza-Yates eds. *Information Retrieval-Data Structures and Algolithms*, chapter4, pp.44-65, Prentice-Hall, Englewood Cliffs, NJ, 1992.