分散ハッシュテーブルにおけるN-gramを用いた部分一致検索の効率化

椚田 博之† 蓑原 隆‡
拓殖大学大学院工学研究科† 拓殖大学工学部‡

1 はじめに

P2P コンテンツ共有システムの形態の 1 つとして，分散ハッシュテーブル（DHT）が提案されている [1]. DHT では，コンテンツのファイル名やキーワードにハッシュ関数を用いて，コンテンツの配置を決定することで，高速に検索を実現している．しかし，登録に使用したキーワードのハッシュ値で検索を行うため，検索キーワードとの完全一致検索しかできないという問題がある．この問題を解決するため，コンテンツの登録時と検索時に短い固定長（N）の文字列（N-gram）をキーとしてコンテンツを登録し，検索キーワードを N 文字に分割したキーの検索結果に対して，キーの連続性を調べる方法 [2] が提案されている．この手法を用いることで，検索キーワードとの部分一致検索が可能である．しかし，論文の抄録検索などの文書データで N-gram に一致するコンテンツの数が増加した場合に本来の検索キーワードとの一致を調べるために中間結果として送られるデータ量が大きくなり大量のトラフィックが発生する可能性がある．

本研究では，DHT 上での N-gram を用いた部分一致検索における中間結果，トラフィックの低減に対してキャッシュによる検索結果の保存，および N.M-gram 法の利用が効果的であると考え，これらの手法について DHT での処理モデルを考え，論文抄録の検索を対象として評価する．

2 DHT 上の N-gram を用いた部分一致検索

DHT 上で N-gram を用いて論文のタイトルと抄録の部分一致検索を行う手法について 2-gram を用いた例を説明する（図 1）．まず，検索対象となるコンテンツ（論文）はタイトルのハッシュ値をキー ID として DHT 上に登録する．また，論文のタイトル，抄録から連続する N 文字として切り出された N-gram について論文を表すキー ID，タイトルと抄録中での N-gram の位置を示す位置情報を N-gram のハッシュ値をキーとして DHT 上に登録する．検索は検索キーワードを分解した N-gram の担当ノード N-gram の連続性を考慮して順次結果をしぼり込むことで行うものとする．例えば，キーワード ”次世代 ”について検索した場合は次のようになる．

1. Hash（次世）担当ノードから，論文 ID，位置情報のリストを中間結果として Hash（世代）担当ノードに送る．
2. Hash（世代）担当ノードは，送られてきた各論文 ID について，その位置情報と自身の位置情報の連続性を調べ結果をしぼり込む．

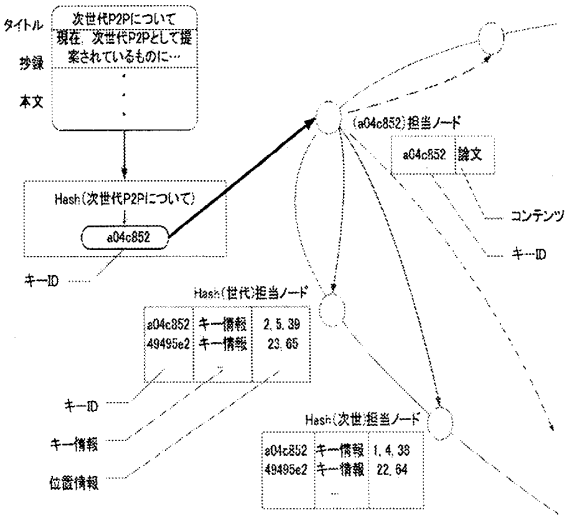

図 1: DHT 上の 2-gram を用いた登録処理

3 検索結果のキャッシュによる効率化

同じキーワードに対する検索時の中間処理を削減するため，検索結果をキャッシュとして検索に使用したキーワードのハッシュ値を担当するノードに登録し，次

Improvement of Substring Queries of N-gram in DHT Network.
†Hiroyuki KUNIGITA, Graduate School of Engineering, Takushoku University
‡Takashi MINOHARA, Department of computer Science, Takushoku University

回以降，同じキーワードで検索した際に登録した検索結果を利用することが考えられる．

例えば，ユーザがキーワード”次世代”について検索を行った場合，まずキーワードのハッシュ値であるHash（次世代）担当ノードへ問い合わせを行う．Hash（次世代）担当ノードに検索結果が登録されている場合は，その結果を返すことで検索処理を終了する．検索結果が登録れていない場合は，キャッシュを行わない場合と同様に，Hash（次世），Hash（世代）担当ノードへ順に問い合わせることで検索を行う．このとき，Hash（世代）担当ノードは検索結果をキーワードのハッシュ値を担当するHash（次世代）担当ノードへ登録をする．キャッシュを用いることで，一度検索結果が保存されたキーワードについての検索では中間結果の通信が不要になる．

キャッシュによる検索結果の保存についてのシミュレーションを行った．シミュレーションに用いるコンテンツとして論文のタイトルと抄録の組を1万本用意した．そして，集めた論文から長さが3以上の名詞の抽出を行い，かつ出現頻度が中頻度のものである780個の名詞からランダムに選択したキーワードで検索を行った．

ハッシュ関数は160ビットの値を返すSHA-1とし，中間結果を求める際に発生する平均トラフィック量を測った（図2）．

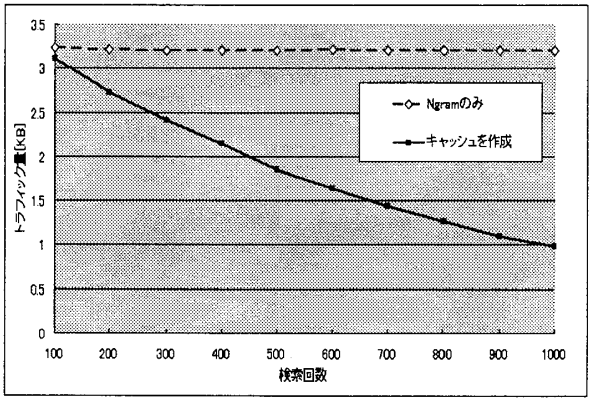

図 2: 検索回数を変化させたときの平均トラフィック量

この結果から，検索回数が比較的少なくキャッシュのヒット率が高くないと考えられる状況においてもトラフィック削減の効果が期待できることがわかる．このシミュレーションでは，キーワードはランダムに選択して行ったが，キーワードに偏りがある場合は，さらにキャッシュの効果はあると考えられる．

4 N.M-gramによる効率化

N-gramによる検索の拡張としてインデックス情報を用いる代わりに後続するM文字のハッシュ値を登録し，N～N+M文字の一致検索を効率化するN.M-gram[3]が提案されている．本研究では，N.M-gramをDHTに利用した場合の効果を考える．N.M-gramでは最初のノードでN+M文字までの一致検索を行うことができるため，N+M文字までの検索では中間結果のトラフィックが発生しない．N+M文字を超える検索についてはN-gramと同様，1文字ずつずらしたキーワードについて順に結果をしぼり込んで行くものとする．このようにした場合，N-gramによる検索と比較して最初のMノード分の中間結果のトラフィックを削減できる．

N.M-gramを用いた場合の中間結果を求める際に発生するトラフィック量を調べるため，前節と同様の環境でシミュレーションを行った．結果より，N.M-gramを用いた場合のトラフィック量は平均0.49KBであった．また，N-gramを用いた場合のトラフィック量は平均3.22KBなので，N.M-gramを用いることで，N-gramを用いたときよりも，トラフィック量を約0.15倍に低減させることができた．

5 おわりに

本論文では，DHT上でのN-gramを用いた部分一致検索における中間結果を求める際に発生するトラフィック低減を目的として検索結果のキャッシュを行う方法，およびN.M-gramを用いる方法について処理モデルを作成し，シミュレーションによって評価した．シミュレーション結果から2つの方法がそれぞれトラフィック削減に効果があることが確認された．

参考文献

[1] Stoica, I., Robert, M., David, K., Frans, K. and Hari, B. : A Scalable Peer-to-Peer Lookup Service for Internet Applications, *ACM SIGCOMM'01*, pp.140-152 August, 2001.

[2] Harren, M., Hellerstein, J., Huebsch, R., Loo, B., Shenker, S. and Stoica, I. : Complex Queries in DHT-Based Peer-to-Peer Networks, *Proc IPTPS02*, pp.242-250 (2002).

[3] 平林 幹雄，江渡 浩一郎：N.M-gram：ハッシュ値付き N-gram 法による転置インデックスの実現，情報処理学会研究報告，Vol.140, No.2 pp.215-222 (2006).