データベースシステム　119－3
（１９９９．７．２１）

距離の公理に基づいた情報放送フィルタリング方式

西山揚子[†*]　大和田俊和[†*]　浅田一繁[†*]　飯沢篤志[†*]　古瀬一隆[‡]
[†]株式会社次世代情報放送システム研究所
*株式会社リコー
[‡]茨城大学工学部情報工学科

要旨

グラフ構造を持つメタ情報を放送で配信・受信するシステムにおける効果的なフィルタリング方式を提案する．本方式では，受信の際，利用者が興味を持つノードと配信されてきたノード間の距離を計算し，利用者が興味を持つノードに近いノードを優先的に受信機に蓄積することによってフィルタリングが実現される．グラフ上の距離計算は高いコストを要するが，本方式では距離の公理に基づいきた近似方式により，受信機における距離計算を簡単にする．この結果，処理能力の低い受信機でも効果的なフィルタリングを行うことが可能となる．

Filtering Mechanism for Broadcasting Information Based on the Triangle Inequality

Yoko Nishiyama[†*]　Toshikazu Ohwada[†*]　Kazushige Asada[†*]　Atsusi Iizawa[†*]　Kazutaka Furuse[‡]
[†]Information Broadcasting Laboratories, Inc.
[*]RICOH, Co.,Ltd.
[‡]Department of Computer and Information Sciences, Ibaraki University

Abstract

In this paper, we propose a mechanism of filtering graph structured meta-data delivered in broadcasting environments. In this method, the distance between nodes in which user is interested and delivered nodes is calculated. The filtering is performed by accumulating delivered nodes near the nodes in which user is interested in each client. Since the cost of calculating distances between nodes is high, we eliminated dynamic distance calculation using approximation based on the triangle inequality. This mechanism performs effective filtering on clients without high processing power.

1.　はじめに

インターネットの普及およびディジタル放送時代の到来に伴い，アクセス可能な情報が急速に増大しつづけている．この結果，膨大な情報の中から必要とする情報を効率

† 株式会社リコーより株式会社次世代情報放送システム研究所へ兼任出向中。
The authors are partly on loan from Ricoh Company, Ltd. to Information Broadcasting Laboratories, Inc.

よく選別することが益々困難になりつつある．このため，個々の情報の概説や所在といったメタ情報を提供し，利用者の効率的な情報選別を支援することが不可欠である．しかし，Web における検索エンジンのようなクライアント・サーバモデルに基づいた方式では，サーバへの処理の集中や通信の負荷といった問題[1][2]があるため，数千万規模のクライアントによる同時使用は実用的でない．

そこで我々は，グラフ構造で表現されたメタ情報を対象とし，そのノードを放送で配信するフレームワークを提案した[3][4][5]．このフレームワークにおいては，数千万規模のクライアントを想定している．デジタル多チャンネル放送における1チャンネルが5Mbpsの伝送速度を持つことを想定する場合，受信側には高速な受信・蓄積処理が必要とされるが，一般家庭に配置される受信機に高い水準の処理能力や記憶容量を求めることはできない．そこで，大規模な情報を高速に処理能力の低い受信機にどのように蓄積させ，クライアントの必要性を満たすかが大きな課題となる．

本論文では，大規模なグラフ構造を持つメタ情報を処理能力の低い受信機でフィルタリングする方式について述べる．まず，2章において，我々のフィルタリングモデルを説明する．次に3章で，フィルタリング処理コストを下げるための工夫について述べる．第4章では本方式における基本要素に関する考察について述べる．最後に本稿での議論および今後の課題についてまとめる．

2. フィルタリングモデル

我々の方式におけるフィルタリングモデルを図1に示す．まず，グラフ構造で表現されたメタ情報のノード集合に対して，距離関数δを定義し，距離空間を構築する．次に，受信の際，利用者が興味を持つノード群を代表するノード(代表点と呼ぶ)hと配信されてきたノード n_1, n_2 間の距離 $\delta(h,n_1)$, $\delta(h,n_2)$を計算し，代表点に近いノード n_2 を利用者にアクセスされる可能性の高いノードとみなして，受信機に蓄積する．

2.1. メタ情報のグラフ表現

本論文では，配信するメタ情報全体はグラフG＝(V, E)で定義される．ここでVはグラフのノード集合であり，個々のメタ情報を表す．E(G)はグラフのリンク集合であり，メタ情報間の関連を表す．リンクに重みを付加することによって、メタ情報間の関連をより厳密に表現できる．

本論文において，グラフの規模は|V(G)|が約百万，|E(G)|が約1千万程度を想定する．個々のノードには，テキ

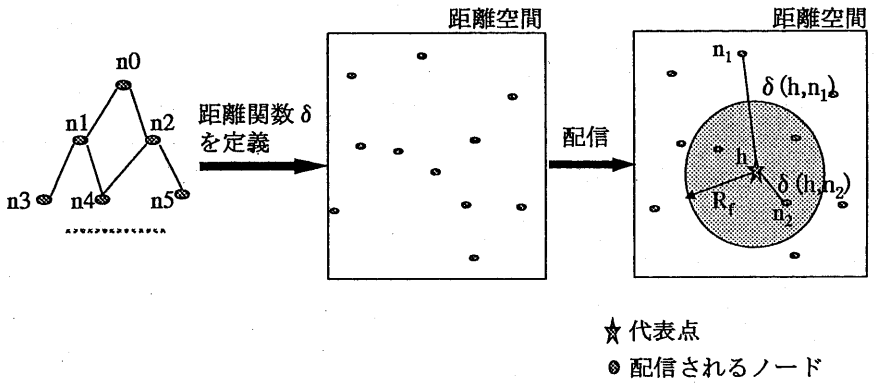

図1 フィルタリングモデル

ストデータだけではなく，画像や音声といったマルチメディア情報も含まれることを考慮し，ノード毎のデータ量は平均100Kbyte程度を想定する．個々のリンクには両端ノードの識別子(8byte)および重み情報(4byte)を含み，そのデータ量は 20byte と仮定する．この仮定に基づくと，E(G)は200Mbyte，V(G)は100Gbyteである．最大1Gbyteの記憶容量を使える受信機を想定している場合，V(G)全体を蓄積することはできないが，E(G)を蓄積するのは可能である．

利用者は，受信機に蓄積されているリンク情報 E(G)を利用して，グラフを辿りながらノードをアクセスする。辿ったノードが受信機に蓄積されていない場合は，通信回線を利用して配信側から取得するという利用形態を想定している．

配信するメタ情報の変更は，グラフGの更新に反映される．更新の受信機への伝播は，次回の放送まで延遅する方式や，動的スケジューリングによってリアルタイムに伝播するという方式が考えられる．

2.2. 距離空間

グラフ構造を持つデータをフィルタリングする際，配信されてきたノードを受信機に蓄積するべきか否かを判断する．このために，個々のメタ情報間の関連を推定する必要がある．本方式では，距離空間に基づいて，メタ情報間の関連を定量的に表現する．配信するメタ情報G=(V,E)に対して，距離空間Ωは下記のように定義する．

$$\Omega=(V(G),\ \delta)$$

ここで，V(G) はグラフのノード集合である．δ は V(G)の直積集合 $V(G)^2=\{(\phi_1,\ \phi_2)|\phi_1\in V(G),\ \phi_2\in V(G)\}$の元 (ϕ_1,ϕ_2)を実数に写像する関数であり，下記の三つの条件を満たす．

$\forall\ \phi_1,\ \phi_2,\ \phi_3\in V(G)$,

(1) $\phi_1=\phi_2 \Leftrightarrow \delta(\phi_1,\ \phi_2)=0$

(2) $\delta(\phi_1,\ \phi_2)=\delta(\phi_2,\ \phi_1)$

(3) $\delta(\phi_1,\ \phi_2)+\delta(\phi_2,\ \phi_3)\geqq\delta(\phi_1,\ \phi_3)$

ここで，(3)は三角不等式(triangle inquality)と呼ばれる式である．距離関数 δ としては，様々なものが考えられるが，後述するようにこの距離はフィルタリングに利用するので，利用者のアクセスパターンを反映したものが望ましい．このようなものとしては，例えばグラフにおける2ノード間の最短経路などが考えられる．距離空間において，$\forall\ \phi_1,\ \phi_2,\ \phi_3\in V(G)$に対して、$\delta(\phi_1,\ \phi_2)<\delta(\phi_1,\ \phi_3)$が成り立つならば、ノード$\phi_1$と$\phi_2$の関連がノード$\phi_1$と$\phi_3$の関連より緊密であると考える。

2.3. 代表点

本方式では，利用者の興味を持つノード群を代表するノードを代表点と呼ぶ．代表点の選定アルゴリズムとしては様々なものが考えられる．例えば，アクセスが集中し，互いに距離が比較的近いノードの集合の中心点を代表点とする方法がある．

一般的に，利用者は複数の興味を持ち，かつ，利用者同士では異なる興味を持つと考えられる．従って，本論文では，個々の利用者の代表点は別々なノードだと考え，代表点は利用者毎に複数(数～数十個)存在すると想定する． さらに，利用者の興味の変化に伴い，代表点も変化すると考えられる．ただし，代表点が変化する頻度は低いと思われる．

2.4. フィルタリング半径

本方式では，代表点に近いノードがアクセスされる確率が高いものと仮定する．具体的には，距離空間において，代表点を中心として，半径R_f 以内のノードを利用者がアクセスする可能性が高いノードとみなし，受信機に蓄積する．つまり，利用者の代表点集合を $H\subset V(G)$とした場合，ノード $v\in V(G)$に対して,$\exists h\in H,\ \delta(v,\ h)<R_f$ならば，v を受信機に蓄積する．この$R_f$をフィルタリング半径と呼ぶ．

フィルタリング半径は代表点毎に規定する．例えば，より頻繁にアクセスされる代表点のフィルタリング半径をより大きく設定することが考えられる．フィルタリング半径が大きければ，受信機に蓄積されるノードの数も増加する傾向にある．受信機のキャッシュが溢れた場合には，キャッシュ置き換えアルゴリズムによって，蓄積されているノードを受信機から削除する．

3. 距離の公理に基づく距離の近似

前述のフィルタリングモデルでは，配信されたメタ情報をノード単位でフィルタリングするので，フィルタリング精度が高い．しかしながら，配信されてきたノードと代表点間の距離 δ (h,n)を動的に求めるのは処理能力の低い受信機にとっては，大きな負担となる．そこで，我々は，(1) 受信機における距離計算を簡単にし，(2) フィルタリング時のディスクアクセスをなくすことを目的として，δ (h,n)を近似的に求める方式を提案する．

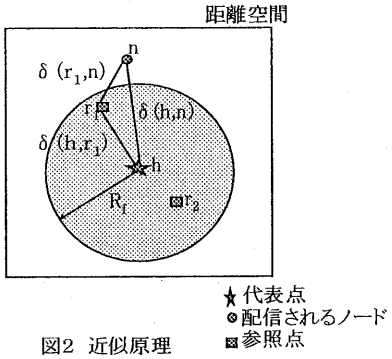

図2 近似原理

3.1. 参照点と近似原理

受信機における距離計算を簡単にするために，参照点という概念を導入する．まず，距離空間において，幾つかのノードr_1, r_2を静的に参照点として選択する(図2)．次に，ノードを放送する前，各ノードと参照点間の距離 δ (r_1, n)を計算し放送する．受信機側は，放送されてきた距離のうち，代表点と参照点間の距離 δ (h, r_1)のみを受信機に蓄積しておく．ノードを配信する際，ノードと参照点間の距離 δ (r_1, n)も一緒に配信する．受信の際，図2に示すように，距離の公理 δ (h, r_1) + δ (r_1, n) ≧ δ (h, n)に基づいて，配信側で計算しておいた δ (r_1, n)と受信機に蓄積されているδ (h, r_1)を用いて，δ (h, n)の近似解を求める．

参照点は配信側で選定するが，参照点の選定アルゴリズムは，本論文では特定しない．

3.2. ノードと参照点間の距離計算

ノードと参照点間の距離計算は，配信の前，配信側で行う．処理能力の低い受信機に比べて，配信側の処理能力には高い水準を期待できる．

しかし，グラフのノードとすべての参照点との間の距離を計算する必要はない．すなわち，ある程度以上離れている参照点は，距離を無限大とみなすことができる．距離計算する際，具体的にどの程度というのは距離計算半径 R_r を

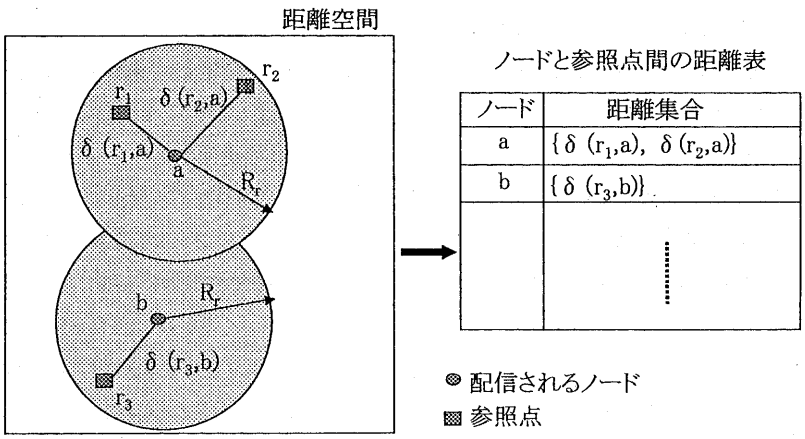

図3 ノードと参考点間の距離計算

用いて規定する．図3には距離計算と距離計算半径の原理が示されているように，距離空間において，任意ノードaを円心として，半径 R_r 以内の参照点r_1，r_2のみを距離計算の対象とする．距離 $\delta(r_1,a)$ と $\delta(r_2,a)$を計算し，距離表に記入する．しかし，ノードaにとって，参照点r_3は半径 R_r 外にあるため，ノード a と参照点r_3 間の距離を計算せずに無限大とみなし，ノードと参照間の距離表には記入しない．この半径 R_rは距離計算半径と呼ぶ．

距離計算半径 R_rは配信側で決める．R_rが小さいほど図3に示すノードと参照点間の距離表のサイズが小さくなり，放送するデータ量を低減できる．

3.3. 代表点と参照点間の距離

フィルタリング時のディスクアクセスをなくすため，$\delta(r,n)$の計算と同様な手法を$\delta(h,r)$の蓄積にも適用する．つまり，図4に示されているように，距離空間において，任意の代表点h_1を中心として，半径R_h以内の参照点r_2とr_3のみを距離蓄積の対象とする．距離 $\delta(h_1, r_2)$ と $\delta(h_1, r_3)$を代表点と参照点間の距離表に蓄積する．しかし，ノード h_1 にとって，参照点 r_4 ，r_5 ，r_6 ，r_7は半径R_h外にあるため，代表点 h_1 とこれらの参照点間の距離を無限大とみなし，ノードと参照点間の距離表には蓄積しない．この半径 R_h を距離蓄積半径と呼ぶ．

これにより，図4に示す代表点と参照点間の距離表のサイズが大幅に圧縮可能となり，受信機のメモリに配置可能となる．この結果，ノードを受信する際，ディスクをアクセスせずに，距離$\delta(h,r)$を得ることができる．

距離計算半径 R_rと違って，距離蓄積半径 R_hは受信側で決める．R_h はフィルタィルタリング R_f を考慮した上で選定するべきである．

4. 考察

ここでは，本方式における距離関数，参照点，代表点の性質について考察する．

4.1. 距離関数について

本論文では、Gが無向グラフか有向グラフかについては仮定しないが、実際に，本方式におけるメタ情報を Web 上のホームページと想定する場合，グラフは有向グラフになるのが自然である．この場合，距離関数 δ を最短経路とすると，条件(2)の対称の公理 $\delta(\phi_1, \phi_2) = \delta(\phi_2, \phi_1)$ が成りたたなくなる．そこで，任意代表点hから任意参照点r間の最短経路を$\delta'(h, r)$をとし，さらに,rから配信された任意のn間の距離を$\delta'(r,n)$とすると，条件(3)の三角不等

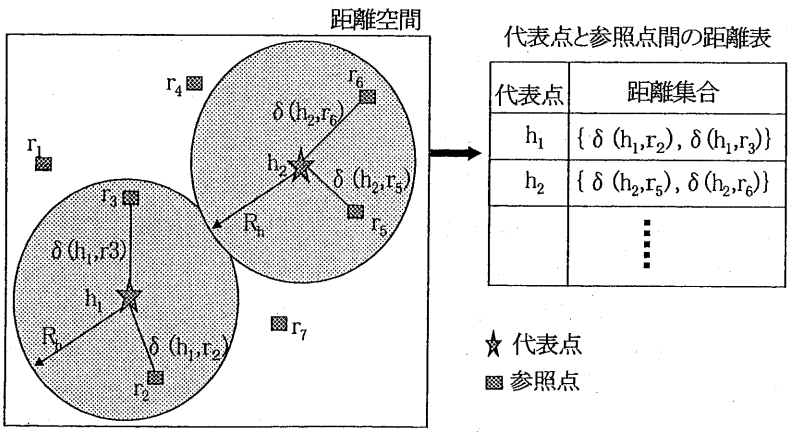

図4. 代表点と参考点間の距離の蓄積

式 $\delta'(h,r)+\delta'(r,n) \geqq \delta'(h,n)$が成り立つので，$\delta'(h,r)+\delta'(r,n)$を$\delta'(h,n)$の近似解として使うことが可能である．

これにより，本論文で提案した近似方式を適用するためには，距離関数δは三つの条件を全て満たす必要がないことがわかる．すなわち，条件(1)および条件(2)の公理さえみたせば，本論文で提案した近似方式が適用可能である．

4.2. 参照点について

一般に，参照点の密度が高ければ，近似精度が向上するが，放送のデータ量も大きくなる．なぜならば，参考点の増加に伴い，ノードと参照点間の距離$\delta(r, n)$の集合も増加するからである．この距離集合はノードと共に配信される．

任意のノードが何れかの参照点を中心とし，R_rを半径とする円に入らないといけない．言いかえれば，図3に示す距離表における任意の距離集合が空集合になってはならない．これは距離集合が空集合であるノードはすべてのクライアントに取りこぼされるからである．この前提を保つために，参照点の密度を大きくするのは有効であるが，前述のように，参照点の密度の増加に伴い，放送データ量も増加する．適切な参照点の選定が，本方式の性能に大きく影響すると考えられる．

4.3. 代表点と参照点間の距離について

本論文では，代表点と参照点間の距離表は配信してきたノードと参照点間の距離の表に基づいて蓄積することを提案しているが，距離のための放送データ量が無視できなくなる場合，代表点と参照点間の距離を受信機側で計算することも考えられる．つまり，配信側は参照点の識別子を事前に配信し，受信側はこれと受信機に蓄積されているグラフのリンク集合 E(G)を利用して，代表点と参照点間の距離を計算しておく．ただし，このため，受信機の処理能力の水準がある程度高くなる．もちろん，適切な距離関数δを選択することによって距離計算コストを削減できるが，最終的には，放送データ量と受信機の処理能力のバランスの問題になる．

5. まとめ

本論文では，グラフ構造を持つメタ情報を放送で配信・受信するシステムにおけるフィルタリング方式を提案した．本方式は，グラフのノード集合に構築される距離空間において，利用者の興味を持つノードに近いノードが高い確率でアクセスされると想定し，受信機に蓄積する．さらに，距離の公理に基づき，距離計算を近似的に行うことによって，受信機におけるフィルタリング処理コストを低減させる．この方式は処理能力の低い受信機でも実装可能という特徴を持つ．

今後は，シミュレーションを行い，本方式の有効性を評価する．さらに，参照点の選定アルゴリズムおよびキャッシング置き換えアルゴリズムと本方式の関連について検討し，本研究を進めていく．

6. 参考文献

[1] T.Imielinski, S.Viswanathan, B.R.Badrinath. "Energy Efficient Indexing on Air", Proc. ACM SIGMOD Conf., Minneapolis, MN, May, 1994.

[2] Swarup Acharya, Rafael Alonso, Michael Franklin, Stanley Zdonik. "Broadcast Disks:Data Management for Asymmetric CommunicationEnvironments", Proc. ACM SIGMOD Conf., SanJose, CA, May, 1995.

[3] 飯沢篤志，浅田一繁，白田由香利：「情報放送のための超大規模分散データベースシステム」，情報処理学会研究報告，97-DBS-113-44, 1997.

[4] 大和田俊和，浅田一繁，飯沢篤志，古瀬一隆:「デジタル放送のためのインデックス情報の断片化方式に関する検討」，情報処理学会研究報告，98-DBS-116-29, 1998.

[5] 大和田俊和，浅田一繁，飯沢篤志，古瀬一隆:「デジタル放送のためのグラフ構造に基づくインデックス情報による検索方式」，情報処理学会研究報告，DEWS'98-6A-1, 1998.