

ユーザ閲覧コストに着目したメタデータ配信ネットワークの基礎検討

馬場 宏基[†] 佐々木 祥[†] 宮田 高道[†] 小林 亜樹^{††} 酒井 善則^{†,††}

[†] 東京工業大学 大学院理工学研究科 〒152-8552 目黒区大岡山 2-12-1

^{††} メディア教育開発センター 〒261-0014 千葉市美浜区若葉 2-12

E-mail: [†]{hbaba,brnw,miyata}@net.ss.titech.ac.jp, ^{††}aki@nime.ac.jp, ^{†††}ys@ss.titech.ac.jp

あらまし コンテンツ流通におけるユーザの検索行動は、1) 検索システムが提示するメタデータの閲覧、2) メタデータの内容に基づくコンテンツ閲覧の判断、3) コンテンツの閲覧、の3つのステップを情報要求に適合するまで繰り返すものであるといえる。このユーザ負担は、メタデータおよびコンテンツの閲覧コストの総和として表現することが出来る。このように分散環境におけるメタデータ配信では、コンテンツ配信とは異なり、ネットワーク負荷ばかりでなくユーザ負担も大きな問題となる。そこで、筆者らはユーザ負担に着目し、コンテンツと分離してメタデータを流通するメタデータ配信ネットワークを提案した。ユーザ負担は提示されるメタデータ順序によって変化し、最適である順序はユーザごとに異なる。そこで筆者らは、まず個々のユーザについての負担を最小化する順序を求めた。さらに、メタデータ配信ネットワークでは、メタデータのノードに対する配置状況と、リンクの伝搬遅延や各ピアでの処理遅延によって決まるノードからの応答順序により、個々のユーザへのメタデータ配信順序を決定することが、ネットワーク負荷の軽減にもつながる。そこで、ユーザが配信された順序そのままメタデータを閲覧するモデルを考え、全ユーザの負担の総和が最小となるメタデータのノードへの配置問題として定式化する。その後、本問題の NP 困難性について議論する。

キーワード メタデータ配信、配置、ユーザコスト、検索、提示順序、MDN

A Proposal of The Metadata Delivery Network focused on User Browsing Cost Reduction

Hiroki BABA[†], Akira SASAKI[†], Takamichi MIYATA[†],

Aki KOBAYASHI^{††}, and Yoshinori SAKAI^{†,††}

[†] Tokyo Institute of Technology, 2-12-1-S3-67, O-okayama, Meguro-ku, Tokyo, 152-8552, Japan

^{††} National Institute of Multimedia Education 2-12, Wakaba, Mihama-ku, Chiba, 261-0014, Japan

E-mail: [†]{hbaba,brnw,miyata}@net.ss.titech.ac.jp, ^{††}aki@nime.ac.jp, ^{†††}ys@ss.titech.ac.jp

Abstract In most retrieval system, user's retrieval procedure consist of the following 3 steps. (1)A user watches metadata provided by the system. (2) If the user wants to see the content corresponding to the metadata , he retrieves the content. (3)If the retrieved content is the target one, the retrieval procedure is terminated, otherwise the user watches the next metadata. These 3 steps are continued until he finds the target. We propose Metadata Delivery Network(MDN) that delivers only metadata to the users. Each metadata has the link to the content, so that users can retrieve corresponding content. Because traffic caused by metadata delivery is small, MDN must be designed to minimize user's retrieving cost until he finds target. We first derive the optimal presentation order of metadata for each user. However the optimal order is not the same for each user, the allocation of metadata in MDN is not easy. We formulate the optimal allocation problem of metadata considering metadata delivery delay, user's cost to watch metadata and to retrieve corresponding content. Finally we discuss the NP-hardness of the problem.

Key words metadata delivery, placement, user cost, retrieval, presentation order, MDN

1. はじめに

現在のコンテンツ配信網 (CDN) では、大量のコンテンツがネットワーク上に散在しているため、未知のコンテンツを取得するために検索が行われる。検索システムは、情報要求に該当するコンテンツを配信するサーバのアドレスやファイル名、およびそのコンテンツに関する特徴となる情報を提示する。これらの情報は、一般的にコンテンツのメタデータと呼ばれる。検索により配信されるメタデータの良し悪しがコンテンツの流通に影響することから、コンテンツの流通促進を目的としてメタデータへの関心が高まっている。そのため、メタデータの生成や配信、流通に関わる研究が盛んに行われている。

文献 [1] は、コンテンツの流通を促進するための環境整備として、一般的な検索・利用・配信のためのメタデータに加えてコンテンツの評価情報を追加したメタデータ体系の提案を行っている。加えて、上記体系にもとづいたメタデータが利用可能なメタデータ交換アーキテクチャおよびデジタルアーカイブシステムの開発を行っている。文献 [2], [3] は、メタデータをハッシュ値を用いて分散して配置し、流通する P2P ネットワークを提案している。さらに、キーワードごとに分散しているメタデータに対して効率的に AND や OR の検索を行う手法を提案しており、計算量やトラフィックといったネットワーク与える負荷に着目した研究である。また、文献 [4] においては、コンテンツを保持するデータベースサーバが持っているフォーマットの異なるメタデータを統合し、異種のデータベースをまとめて検索することが可能となる P2P ネットワークを提案している。メタデータの生成手法に関する検討もなされており、文献 [5] はユーザによるメタデータ生成を助けるインターフェースの提案を行っている。

検索とは、ユーザが検索要求をクエリとしてシステムに入力し、システムより返されたメタデータの中から検索要求に合致するものを選び、目的とするコンテンツを発見することで終了する行動である。各メタデータについて、ユーザはその内容を確認し、そのメタデータが指し示すコンテンツを取得するかどうかの判断を行う。しかしながら、メタデータの内容だけでは、コンテンツが検索の意図に合致するかどうかを正確に判断することができない。このため、目的のコンテンツを表すメタデータを閲覧しても、目的のコンテンツへと到達できなかったり、逆にメタデータの内容から目的のコンテンツであると判断しても、実際のコンテンツは目的に合致しないことがある。したがって、ユーザは、システムから提示されたメタデータ集合の個々のメタデータについて

(1) メタデータの内容確認

(2) メタデータの指し示すコンテンツの内容確認

の 1 だけ、または 1 と 2 の行動を繰り返す。一方、検索システムはメタデータを順序を持ってユーザに提示しており、ユーザはこの順序に従って内容を確認する。以上より、目的のコンテンツに到達するまでに、ユーザは異なるメタデータ、およびコ

ンテンツを多数閲覧することを強いられる。この繰り返しのために、ユーザは時間や手間などを負担することになる。個々のメタデータを見てもユーザがそのメタデータが示すコンテンツが目的に合致するかどうかの正確な判断を下せないため、検索システムがメタデータを提示する順序が重要となる。

そこで、本稿では目的コンテンツに到達するまでのユーザの負担を少なくするために、ユーザへメタデータを提示する順序 (以下、メタデータ提示順序と定義する) を検討する。筆者らは、ユーザの検索コストが少なくすむような順序でメタデータを提示することを目的として、検索システムにおいて、クエリを受け取り、クエリに該当するメタデータの配信順序に注目した概念であるメタデータ配信ネットワーク (Metadata Delivery Network, MDN) を提案している [6], [7]。MDN の背景や特徴については、2 章にて述べる。

本稿では、MDN の最適化に関する基礎検討として、最適なメタデータ提示順序について 3 章にて検討する。まず、ユーザの検索行動をモデル化し、ユーザのメタデータを利用する確率、コンテンツが正解である確率により行動が決まる確率モデルとして表現する。次に、これらの確率を既知として、あるクエリにヒットしたメタデータ集合を順番に閲覧していくときに、検索終了までの負担が最小となるメタデータ提示順序 (最適提示順序) を求める。また、複数のネットワークノードにより構成される分散システム上において最適なメタデータ提示順序を実現する手法の一つとして、メタデータの最適配置を求める。4 章では、ユーザの嗜好などの情報に関しては既知として、MDN の内部でのネットワークノードへのメタデータの配置を最適化する問題を数学的に定義し、この問題の NP 困難性について議論する。

2. メタデータ配信ネットワーク (MDN)

本節では、検索を行うシステムにおいて、ユーザからの検索クエリを受け取り、クエリに該当するメタデータのある順序に従って配信する機能を取り出したモデルとして、メタデータ配信ネットワークを提案する。

2.1 検索におけるメタデータの役割

大量のコンテンツが存在し、情報要求に見合うコンテンツが存在するか否かも分からない状況下では、検索は必須といえる。種々の検索システムにおいて検索結果としてユーザに提供される、タイトルや著者名、要約文といった情報は、コンテンツのメタデータである。ユーザは、このメタデータを閲覧し、内容を確認して実際のコンテンツを閲覧するかどうかを決断した上で、コンテンツの取得、閲覧し内容を確認する。すなわち、コンテンツ閲覧までのユーザ行動は、

(1) クエリに対して検索システムが提示するメタデータの閲覧

(2) メタデータの内容に基づくコンテンツ閲覧の判断

(3) コンテンツの閲覧

の 3 ステップから成り、目的コンテンツに達するまで繰り返さ

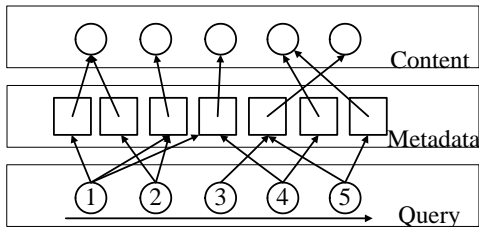


図 1 検索の階層性

れる。ステップ (2) では、ステップ (1) にて閲覧したメタデータに基づいてコンテンツを閲覧するかどうかの判断を行う。これにより、情報要求に合致しないコンテンツの取得と閲覧による手間を省くことができ、単に (1), (3) を繰り返すよりも効率的に目的コンテンツの閲覧に到達できる。以下では、メタデータおよびコンテンツを閲覧してその内容を確認することを単に「閲覧」呼ぶ。

この過程における、クエリ、メタデータ、コンテンツの関係性をクエリの投入順序を考慮して図示したのが図 1 である。一般にユーザは目的コンテンツの閲覧までに、コンテンツよりも判断コストの小さいメタデータを多数閲覧し判断する。したがって、メタデータには、判断のための信頼に足るデータであること、また、よりメタデータの先にあるコンテンツがユーザの情報要求に合致する可能性が高く、伴う負担の小さいメタデータから提示されることが求められる。多彩な嗜好を持ち、多様な情報要求を発するユーザに信頼に足るメタデータを提供するためには、従来より検索に用いられてきたメタデータだけでなく、コンテンツを閲覧したユーザによる評価などのメタデータも必要である。

また、メタデータを適切な順序で提示することで、ユーザの検索負担の低減に直結する。しかし、分散されたネットワークノードにメタデータが格納されているとすると、ネットワークやノードの状況に依存した自然な配信順序でユーザに提示したメタデータ群が、適切な順序であるとは限らない。また、クライアント端末上などで順序の入れ換えを行うのは、応答時間が延びるなどの弊害がある。一方、図 1 からは、メタデータはコンテンツと分離独立して存在し、配信が可能であることが示唆される。そこで、

- CDN の種類を問わずにメタデータを集約、収集し、
- クライアント端末上ではなく、ネットワーク上で自然とユーザの負担を最小化が可能

となる、メタデータ配信だけを行うメタデータ配信ネットワーク (MDN) を提案する。

2.2 MDN の特徴

ユーザの検索のコストは、クエリに対して提示するメタデータの順序が重要である。そこで本稿では、ユーザの検索コストを低減するような順序でメタデータを配信する検索システムを、分散ネットワーク上において実現するものを MDN と呼ぶ。

MDN の特徴として、メタデータを CDN に関係なく集めることで、コンテンツの配信形態とは独立した検索機能の提供で

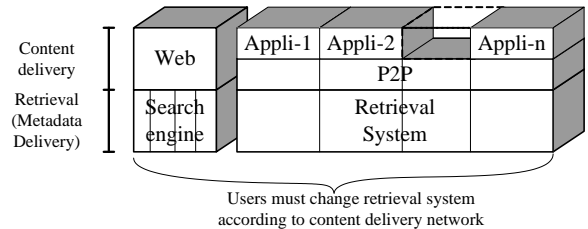


図 2 従来の検索とコンテンツ配信の関係

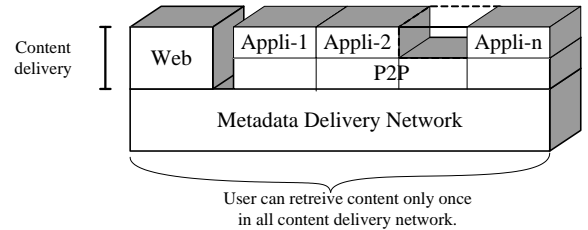


図 3 MDN とコンテンツ配信との関係

きる。図 2 に示すように、従来は CDN ごとに検索システムが存在していた。つまり、WWW の検索には検索エンジンが用いられるが、P2P 等の別の CDN 上のコンテンツの検索には、別の検索手段が必要となる。P2P ファイル共有アプリケーションは多数存在するがそこで提供される検索システムの検索対象は同じ網上のコンテンツについてのみである。このためユーザは複数の CDN に対して、別に検索を行う必要がある。

MDN と各 CDN の関係は、図 3 のようになるため、ユーザは一回のクエリ投入で、複数の CDN 上の検索を完了させることができる。これは、CDN とは独立にメタデータを管理するため、複数の CDN からメタデータを集約するが可能であることによる。また、分散型のネットワークを用いていることにより、容易にメタデータのアップロードが可能となる。これにより、従来より利用されているメタデータ (タイトル、著者、ファイル名) だけでなく、ユーザが閲覧したコンテンツについての感想や要約、タグといった情報をメタデータとして MDN へとアップロード可能とする。この様子を図 4 に示す。ユーザが生成するメタデータの利用により多彩なユーザの嗜好性にあった検索が可能となる。

以上に述べたように、MDN により CDN を問わず検索が可能になり、ユーザのメタデータアップロードを許すことで、より多様な検索要求にこたえることができる。MDN の実現のためには多くの課題が存在するが、本稿では検索時のユーザのコストを最小化する順序をメタデータの最適提示順序とよび

- メタデータの最適提示順序の導出
- 最適提示順序を実現するネットワーク上でのメタデータの配置問題

についての考察を行う。

3. 単一ユーザに対するメタデータ配信の最適化

本章では、MDN により配信されるメタデータの提示順序の評価のため、1 ユーザについてメタデータの提示順序を最適化

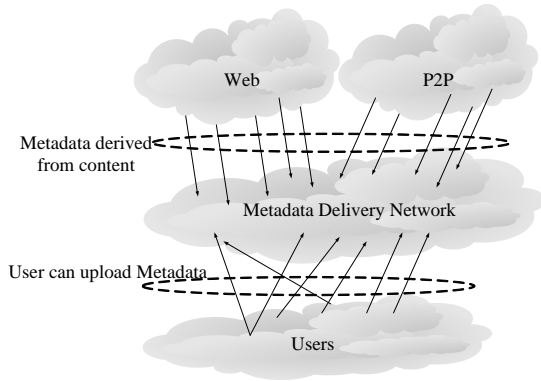


図 4 MDN によるメタデータの収集

する手法について検討する。

検索結果は、クエリにヒットしたメタデータ集合が、ネットワーク上のノードとの距離や、ノード内での並べ替えにより決定される順序に従って並べ替えられる。このため、ユーザの検索の負担が最小となるように最適化する順序を、ユーザの検索行動を確率モデルとして表現して、数学的に導出する。そこで、検索の負担をコストとして定義し、ある順序でメタデータを閲覧することによって生じるコストの期待値を求める。最後に、この値を最小化するメタデータの提示順序（以降、最適提示順序と呼ぶ）を導出する。

3.1 ユーザの検索行動のモデル化

検索において、ユーザは投入したクエリに該当するメタデータの集合をシステムから提示される。メタデータの閲覧は、提示された順序に従って行われる。あるメタデータが検索意図に合致するコンテンツをさしていると判断することを以下では、メタデータを選択と呼ぶ。メタデータを選択した場合にだけ、ユーザはコンテンツを取得、閲覧する。また、ユーザはメタデータを選択しない場合や、選択したにも関わらずコンテンツの内容が目的に合致しない場合には、再度別のメタデータを閲覧して選択の判断を行う。

したがって、ユーザのクエリ投入後の行動は、メタデータの閲覧と選択、コンテンツ閲覧の行動を繰り返すものであるといえる。すなわち、ユーザの検索行動は以下のように五つのステップにモデル化できる。

- (1) クエリをネットワークへと投入する。
- (2) 提示された順序にメタデータの一つを閲覧する。
- (3) メタデータを選択する場合は(4)へ、選択しない場合は(2)へと戻る。
- (4) 当該コンテンツを取得し、閲覧する。
- (5) コンテンツが目的に見合うものであれば検索終了。それ以外なら(2)へと戻る。

各ステップは、図5のフローチャートの括弧内の番号と対応する。

上記の(3)において、メタデータ m を選択する確率をメタ

データごとに $P(m)$ と定義する。また、(5)においてメタデータの選択により得られるコンテンツがユーザの検索意図と合致する確率を、メタデータ m を選択したという条件付きの確率としてコンテンツ c の正解確率 $P(c|m)$ と定義する。

3.2 コストのモデル化

ユーザが検索を行うとき、コンテンツやメタデータの通信によるトラフィックの負荷や、ユーザがそれらを閲覧するために消費する時間等がコストとなる。本稿では、ユーザの負担するコストに着目しており、ネットワーク中での通信コストは考慮しないものとする。メタデータ m を閲覧するコスト $C(m)$ は前節の検索行動モデルにおいて(2)で発生し、コンテンツ c を閲覧するコスト $C(c)$ は(4)で発生する。メタデータを閲覧して選択しない場合や、コンテンツが検索意図に合致しない時には再度(2)へと戻るため、ユーザは上記の検索行動モデルを複数回繰り返すことになる。したがって、検索終了までのユーザの検索コストは、閲覧したメタデータやコンテンツによるコストの総和になる。

3.3 コストの期待値

以上より定義した $P(m)$, $P(c|m)$, $C(m)$, $C(c)$ を用いることで、メタデータのある順序でユーザに提示することによるコストの期待値が導出できる。1章でも述べたように、評価のための最適値導出を行うため、ここでは $P(m)$, $P(c|m)$, $C(m)$, $C(c)$ は既知であると仮定する。

あるクエリに対して得られた n 個のメタデータのうち、 i 番目に提示されるメタデータを $m(i)$ とする。また、ここではメタデータとコンテンツが同数存在し、それらは一対一に対応しているものとする。すなわち、 i 番目のコンテンツ $c(i)$ はユーザが $m(i)$ を閲覧した後に閲覧されるものである。

次に、 i 番目のメタデータおよびコンテンツによって検索が成功する確率と、そのときのコストを $P_s(i)$, $C_s(i)$ 、失敗する確率と、そのときのコストを $P_f(i)$, $C_f(i)$ とすると、これらは以下の式で表される。

$$P_s(i) = P(m(i))P(c(i)|m(i)) \quad (1)$$

$$P_f(i) = 1 - P_s(i) \quad (2)$$

$$C_s(i) = C(m(i)) + C(c(i)) \quad (3)$$

$$C_f(i) = C(m(i)) + \frac{C(c(i))P(m(i))(1 - P(c(i)|m(i)))}{P_f(i)} \quad (4)$$

以上の記号を利用することで、ユーザが k 個目で検索が成功する確率 $P_T(k)$ は、以下のようにあらわされる。

$$\begin{aligned} P_T(k) &= P_f(1)P_f(2) \cdots P_f(k-1)P_s(k) \\ &= \left(\prod_{i=1}^{k-1} P_f(i) \right) P_s(k) \end{aligned} \quad (5)$$

また、その際のコストは以下のようにあらわされる。

$$\begin{aligned} C_T(k) &= C_f(1) + C_f(2) + \cdots + C_f(k-1) + C_s(k) \\ &= C_s(k) + \sum_{j=1}^{k-1} C_f(j) \end{aligned} \quad (6)$$

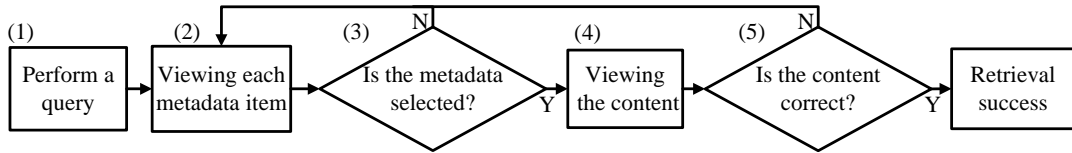


図 5 ユーザの検索行動モデル

したがって、ある順序で n 個のメタデータを提示することによるユーザのコストの期待値 C_{total} は、以下のようにあらわせる。

$$C_{total} = C_T(1)P_T(1) + C_T(2)P_T(2) + \dots + C_T(n)P_T(n)$$

$$= \sum_{l=1}^n C_T(l)P_T(l) \quad (7)$$

以上から、 $P(m), P(c|m), C(m), C(c)$ が既知であるとき、メタデータ m_1, \dots, m_n の順序 $(m(1), \dots, m(n))$ から C_{total} を求める関数が存在することが分かる。これを以降では g と呼ぶ。

$$C_{total} = g((m(1), \dots, m(n))) \quad (8)$$

3.4 最適提示順序

ある順序でメタデータを提示することによるコストの期待値を前節にて導出した。これを用いることで、ネットワーク上で探索コストの異なる情報の最適探索順問題 [8] と同様の議論が可能となり、全てのメタデータ $m_i (i = 1, \dots, n)$ について $\frac{C(m_i)}{P(m_i)P(c_i|m_i)} + \frac{C(c_i)}{P(c_i|m_i)}$ を計算し、この値の昇順にメタデータを提示することが最適提示順序となることが証明できる。具体的な証明は付録に示す。

以上により、1 ユーザ u の検索コストを最小化するメタデータの提示順序が求まった。

4. 複数ユーザに対するメタデータ配信の最適化

本章では、MDN 上でメタデータの提示順序の最適化の基礎検討として、各ノードに対するメタデータ配置の最適化を検討する。そこで、数学的に定式化可能な条件におけるメタデータの最適な配置を求める。

個々のノードには、ユーザが接続しているものとし、クエリはネットワークトポロジに沿ってノード間を転送されるため、ユーザに対するメタデータの配信順序はメタデータ配置により決定される。しかし、ネットワークトポロジと各ノードに配置可能なメタデータ数の制約により、すべてのユーザについて最適提示順序を実現することが困難である。そこで、ユーザ全体のコストの総和を最小化するメタデータ配置を最適配置と定義する。この配置を求める問題であるメタデータ最適配置問題を定式化し、本問題の最適解の導出について議論する。

4.1 メタデータの提示順序

本節では、MDN 上の検索によるメタデータの提示順序の決定要因について述べる。

各ノードでの検索結果がユーザに配信されるまでの遅延により、図 6 に示すようにノード間で検索結果をユーザに返す順序が存在する。これは、クエリと検索結果の転送にかかる時間、クエリとメタデータとのマッチングにかかる遅延がノード

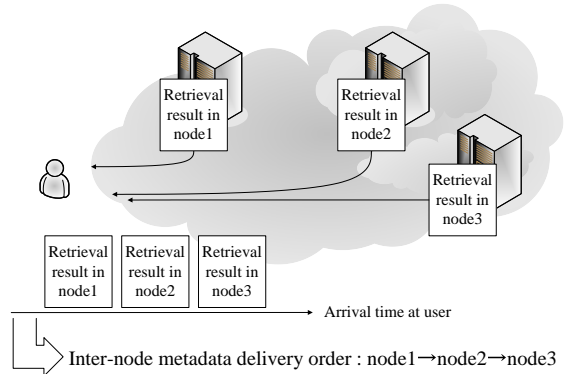


図 6 ノード間配信順序:各ノードからの検索結果がユーザに届く順序

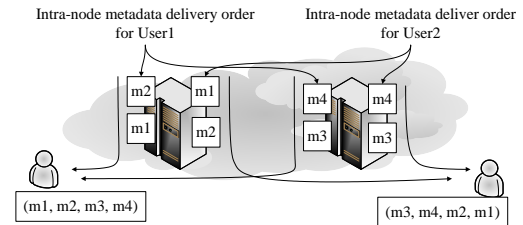


図 7 ノード内配信順序:同一ノード内でのメタデータが配信される順序

ごとに異なることによる。以降は、この順序をユーザに対する「ノード間配信順序」と呼ぶ。

また、各ノードでの検索結果も、ユーザが目にする順序が存在するため、図 7 に示すように個々のノードが返す検索結果内でメタデータ間に順序が存在する。以降では、この順序を「ノード内配信順序」と呼ぶ。ノード内配信順序は、仮定から最適提示順序に従うものとする。

4.2 メタデータ最適配置問題の定式化

3 章にて示したように、ユーザに対してメタデータを見せる順序から検索コストの期待値が導出可能である。各ノードに配置されたメタデータは、ノード内配信順序に従ってユーザに配信される。また、各ノードからのメタデータ列を、ノード間配信順序をもとに並べることで、ユーザへのメタデータ提示順序が決定する。以上の関係により、メタデータの配置からすべてのユーザについてメタデータの提示順序が求まるため、各ユーザの検索コストの期待値を導出することができる。最後に、その和を求めることで全ユーザのコストの期待値の総和が求まる。これを最小化する配置を求めるのがメタデータ最適配置問題である。

以下では、メタデータ最適配置問題の定式化を行う。前提として、ノード間配信順序は既知とする。また、すべてのユーザについて $P(m), P(c|m), C(m), C(c)$ が既知であると仮定し、

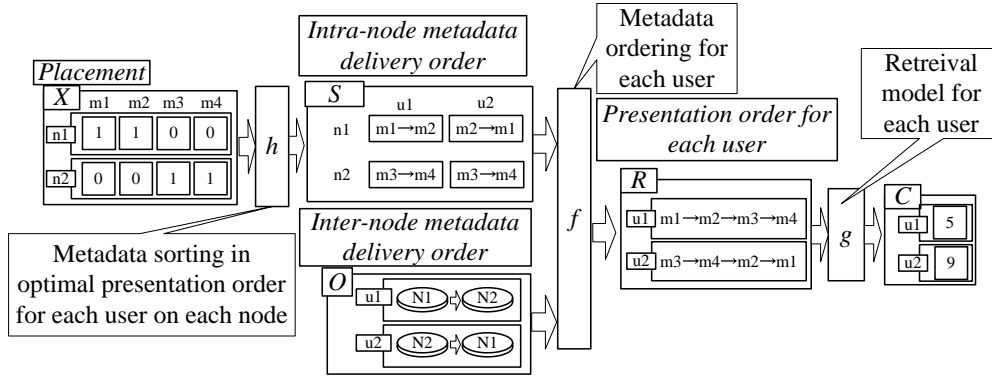


図 8 X, S, O, R, C の関係

ノード内配信順序が計算可能であるとする。

まず、ノード間配信順序を以下のように表現する。

- $O = \{o_{ki}\}$: ユーザ u_k からのクエリに対するノード n_i のノード間配信順序

$o_{ki} = 2$ は、ユーザ u_k に対してノード n_i からの検索結果が 2 番目に配信されることを意味する。

次に、メタデータの配置状態を以下のように表現する。

- $X = \{x_{ij}\}$: n_i におけるメタデータ m_j の配置状態 ($x_{ij} \in \{0, 1\}$)

$x_{ij} = 1$ は、 n_i にメタデータ m_j が配置されている状態を示す。逆に、 $x_{ij} = 0$ ならば配置されていない状態を示す。

ここから、各ユーザへの配信順序を定式化していく。まず、配置状態 X からノード n_i に配置されたメタデータ集合 M_i がわかる。 M_i の u_k に対するノード内配信順序 s_{ik} は、最適提示順序に従う。ここで、各ユーザへのノード内配信順序行列 S を以下のように定義する。

- $S = \{s_{ik}\}$: n_i から u_k に対して配信されるメタデータ列

また、メタデータ配置 X から n_i に配置されたメタデータ集合を取り出し、 u_k の最適提示順序に並べ替えることを、すべてのユーザとノードの組み合わせに対して行う関数を h と定義する。 h により S は以下のように求まる。

$$S = h(X) \quad (9)$$

これは、図 7 においては

$$\begin{aligned} M_1 &= \{m1, m2\}, & M_2 &= \{m3, m4\} \\ s_{11} &= (m1, m2), & s_{12} &= (m2, m1), \\ s_{21} &= (m3, m4), & s_{22} &= (m3, m4) \end{aligned} \quad (10)$$

となることと対応する。

次に、メタデータ提示順序 R を以下のように表現する。

- $R = \{r_{kj}\}$: u_k に対して j 番目に提示するメタデータ ($0 < r_{kj} \leq |M|$)

ここで、 O に従って S の要素を並べることにより、ユーザへのメタデータ提示順序 R が求まる。よって、 O と S からメタデータ提示順序 R を求めることができる関数 f が存在する。

$$R = f(O, S) \quad (11)$$

これは、図 7 では式 (10) で求めた順序 $s_{11}, s_{12}, s_{21}, s_{22}$ を O の並べることを意味する。したがって、この場合の R は以下のように求まる。

$$\begin{aligned} O &= \begin{pmatrix} 1 & 2 \\ 2 & 1 \end{pmatrix} \text{ のとき} \\ R &= \begin{pmatrix} (s_{11}, s_{21}) \\ (s_{22}, s_{12}) \end{pmatrix} \\ &= \begin{pmatrix} (m1, m2, m3, m4) \\ (m3, m4, m2, m1) \end{pmatrix} \end{aligned}$$

以上により、ノード間配信順序 O とノード内配信順序 S からメタデータの提示順序 R が求まった。そこで、式 (8) を用いて各ユーザのコストの期待値 c_k が求まる。

- $C = \{c_k\}$: u_k の総コストの期待値

X, S, O, R, C の具体例を図 8 に示す。以上より、メタデータ最適配置問題は以下の条件を満たす X を求める問題として定式化される。

$$\min : |C| = \sum_i c_i = |g(f(O, h(X, u_i)))| \quad (12)$$

$$\text{s.t.} : \forall i \sum_j x_{ij} \leq \text{Cap}_i \quad (13)$$

$$\forall i, \forall j \ x_{ij} \in \{0, 1\} \quad (14)$$

$$\forall j \sum_i x_{ij} \geq 1 \quad (15)$$

式 (15) は、各メタデータが 1 つ以上存在することを示す。ただし、 Cap_i は、ノード n_i に配置できるメタデータの最大数であり、 $|\cdot|$ はベクトルの要素の和を表すものとする。この問題を解くことにより、全ユーザについてのコストの総和を最小とする最適なメタデータの配置を求めることができる。

4.3 NP 困難性の検討

前節にて定式化したメタデータの最適配置問題を満たす X を求めることで、MDN 上でのメタデータ配置の最適化が可能となる。本節では、この問題の NP 困難性について検討する。

本問題では、 m_j を n_i に配置することによって、すべてのユーザについてそのメタデータを見る順序はおおよそ決まる。しかし、その順序でメタデータを見ることによって生じるコストはそれ以前に閲覧するメタデータにより決まるため、 $x_{ij} = 1$ により生じるコストは x_{ij} だけでは定まらない。このことが、最適解の導出を困難としている。

ここで、一般的に NP 困難であることが知られている一般化割当問題 [9] と、本問題との比較を行う。一般化割当問題の詳細は付録にて紹介する。本問題において $x_{ij} = 1$ となることで生じるコストが c'_{ij} と他のメタデータ配置に依存せずに決まる別の問題は、一般化割当問題と同じ問題となる。この別問題は NP 困難であることがわかるが、この仮定による本問題からの計算困難性の増減については解明できていない。しかし、 $x_{ij} = 1$ であることによるコストがそのほかの x_{kl} に依存することで問題は複雑になっており、一般化割当問題と同等、もしくはそれ以上に最適解の導出は困難と予想される。よって、 X の最適化にはヒューリスティックな手法を用いる他ないといえる。

5. まとめ

本稿では、ユーザからクエリを受け取り、クエリに該当するメタデータの配信順序に注目して、分散ネットワークに存在するコンテンツ検索システムをモデル化した、MDN を提案した。このため、MDN は、検索クエリに対してユーザが効率的に検索ができるメタデータ提示順序を実現を目的とする。

MDN が配信するメタデータの配信順序の評価を目的として、個々のユーザのコストを最小化するメタデータの最適提示順序を導出した。これにより、 $P(m), P(m|c), C(m), C(c)$ をユーザごとに既知と仮定したときに、 $\frac{C(m)}{P(m)P(c|m)} + \frac{C(c)}{P(c|m)}$ の値の昇順にメタデータを提示することが最適提示順序であることを示された。次に、MDN 上でメタデータの提示順序の最適化を行うための基礎検討として、トポロジ固定のネットワークにおける各ノードに対するメタデータ配置の最適化について検討した。まず、メタデータの配置を変数として、各ユーザに対するメタデータの提示順序を求め、全ユーザについてコストの期待値を求める関数を導出した。さらに、全ユーザのコストの総和を最小化する問題としてメタデータ最適配置問題を定義し、この問題の NP 困難性を一般化最適問題との類似性から検討した。

今後の検討課題として、一般化割当問題を本問題に還元することによる NP 困難性の証明、およびメタデータ最適配置問題をヒューリスティックに解く手法の検討、および今回は既知と仮定した $P(m), P(m|c), C(m), C(c)$ をユーザの行動から予測する手法の検討があげられる。

文 献

- [1] 藤井 寛, 杉山 武史, 木谷 靖, 岩淵 明, 中原 美沙, 小池 真由美, 藤生 宏, 曾根原 登. “デジタルシネマ流通促進のためのメタデータに関する考察,” 情報処理学会研究報告, 2005-AVM-48, pp. 67-72, 2005.

- [2] T. Suel, C. Mathur, J. Wu, J. Zhang, A. Delis, M. Kharrazi, X. Long and K. Shanmugasundaram. “ODISSEA: A Peer-to-Peer Architecture for Scalable Web Search and Information Retrieval,” in *Proc. of the International Workshop on the Web and Databases (WebDB)*, Jun. 2003.
- [3] M. Cai and M. Frank. “RDFPeers: A Scalable Distributed RDF Repository based on A Structured Peer-to-Peer Network,” in *Proc. of the 12th International World Wide Web Conference (WWW2004)*, pp.650-657, May. 2004.
- [4] W. Nejdl, B. Wolf, C. Qu, S. Decker, M. Sintek, A. Naeve, M. Nilsson, M. Palmer and T. Risch “EDUTELLA: A P2P Networking Infrastructure Based on RDF,” in *Proc. of ACM the 11th World Wide Web Conference (WWW2002)*, pp. 604-615, May 2002.
- [5] K. Cardinaels, M. Meire and E. Duval, “Automating Metadata Generation: the Simple Indexing Interface” in *Proc. of the 14th International World Wide Web Conference 2005 (WWW2005)*, pp. 548-556, May 2005.
- [6] 馬場 宏基, 佐々木 祥, 宮田 高道, 山岡 克式, 酒井 善則. “分散ネットワーク環境におけるメタデータ配信方式 (1) - 検索行動モデル化によるメタデータ配置の検討-,” 電子情報通信学会ソサイエティ大会, pp. 152, B-7-92, 2006.
- [7] 宮田 高道, 馬場 宏基, 小林 亜樹, 酒井 善則. “分散ネットワーク環境におけるメタデータ配信方式 (2) - 最適提示順序の導出-,” 電子情報通信学会ソサイエティ大会, pp. 153, B-7-93, 2006.
- [8] 菅原 真司, 山岡 克式, 酒井 善則, “ネットワークにおける画像情報の効率的探索法に関する検討,” *IEICE Trans. on Commun.*, vol.J81-B-I, no. 8, pp. 484-493, Aug. 1998.
- [9] M. Yagiura, T. Ibaraki and F. Glover, “A path relinking approach for the generalized assignment problem,” in *Proc. of the International Symposium on Scheduling*, pp. 105-108, Jun. 2002.

付 録

最適提示順序の証明

<証明>

メタデータを $m(1), m(2), \dots, m(n)$ の順に閲覧する場合のコストの期待値を C_t とし、 k 番目と $k-1$ 番目のメタデータを入れ替えた場合のコストの期待値を C'_t とすると、

$$C_t - C'_t = (P_T(k-1)C_T(k-1) + P_T(k)C_T(k)) - \left[\frac{P_T(k)}{P(k-1)} (C_T(k) - C(k-1)) + P_T(k-1)P(k)(C_T(k-1) + C(k)) \right] \quad (A.1)$$

となる。全体を $\prod_{i=1}^{k-2} P(i)$ で割ったものを ΔC とすると、

$$\Delta C = \frac{C_f(k-1)}{P_s(k-1)} + (C_s(k-1) - C_f(k-1)) - \left[\frac{C_f(k)}{P_s(k)} + (C_s(k) - C_f(k-1)) \right] \quad (A.2)$$

$$= \frac{C(m(k-1))}{P(m(k-1))P(c(k-1)|m(k-1))} + \frac{C(c(k-1))}{P(c(k-1)|m(k-1))} - \left[\frac{C(m(k))}{P(m(k))P(c(k)|m(k))} + \frac{C(c(k))}{P(c(k))} \right] \quad (A.3)$$

が得られる。(A.3) 式より明らかに、

$$\begin{aligned} & \frac{C(m(k-1))}{P(m(k-1))P(c(k-1)|m(k-1))} + \frac{C(c(k-1))}{P(c(k-1)|m(k-1))} \\ & < \frac{C(m(k))}{P(m(k))P(c(k)|m(k))} + \frac{C(c(k))}{P(c(k)|m(k))} \end{aligned} \quad (\text{A.4})$$

のとき, $C'_i > C_i$ となる.

したがって, 二つのメタデータが $\frac{C(m(i))}{P(m(i))P(c(i)|m(i))} + \frac{C(c(i))}{P(c(i)|m(i))}$ の小さい順序で提示された場合のコストの期待値は, 提示がその逆順で行われた場合と比較して常に小さくなる. このことから, コストの期待値を小さくするためには $\frac{C(m(i))}{P(m(i))P(c(i)|m(i))} + \frac{C(c(i))}{P(c(i)|m(i))}$ の昇順にメタデータの提示を行えばよいことは明らかである.

< 証明終わり >

一般化割当問題 (Generalized Assignment Problem, GAP)

GAP は, 与えられた n 個の仕事 $J = 1, \dots, n$ を m 個のエージェント $I = 1, \dots, m$ に割り当てたとき, 割り当てにともなうコストの総和を最小化する問題である. 仕事 $j \in J$ をエージェント $i \in I$ に割り当てたときのコスト c_{ij} と資源の要求量 a_{ij} , および各エージェント $i \in I$ の利用可能資源量 b_i が与えられている. 各仕事は必ずいずれか 1 つのエージェントに割り当てなければならない. また, 各エージェントに割り当てられた仕事の資源要求量の総和は, そのエージェントの利用可能資源量を超えてはならない. 仕事のエージェントへの割当ては, 写像 $\sigma: J \rightarrow I$ で与えられる. $\sigma(j) = i$ は, 仕事 j をエージェント i に割り当てることを意味する. 便宜上, 0-1 変数 x_{ij} を

$$x_{ij} = 1 \iff \sigma(j) = i$$

と定義する. これを用いると, GAP は以下のように定義できる.

$$\min : \text{cost}(\sigma) = \sum_{i \in I} \sum_{j \in J} c_{ij} x_{ij}$$

$$\text{s.t.} : \forall i \in I, \sum_{j \in J} a_{ij} x_{ij} \leq b_i \quad (\text{A.5})$$

$$\forall j \in J, \sum_{i \in I} x_{ij} = 1 \quad (\text{A.6})$$

$$\forall i \in I \text{ and } \forall j \in J, x_{ij} \in \{1, 0\} \quad (\text{A.7})$$