令和元年度卒業研究論文

URL の情報指向型クラシフィケーション

2020 年 2 月 7 日 (金)

指導教員　井上一成　教授

明石工業高等専門学校
電気情報工学科

報告者　E1533　西 総一朗

目次

第 1 章

序論

1.1 TCP/IP の課題

1983 年から今日のインターネットと呼ばれているネットワークにおいて通信プロトコル TCP/IP がデファクトスタンダードとなった [1]．約 20 年前のインターネットのトラヒックや利用形態は現在とは大きく異なっている．1992 年の全世界のインターネットトラフィックは 1 日あたり約 100 GB であったが, その 10 年後の 2002 年には 1 秒あたり 100 GB に増え，2017 年には 1 秒あたり 45,000 GB 以上に到達した．また利用形態も 2017 年においてはトラヒックの 75% をビデオコンテンツが占めている．Cisco によると全世界のインターネットトラヒックは 2022 年には 150,700 GB/秒となりその 82% をビデオコンテンツが占めると予測されている [2]．

また，インターネットの使用目的も変遷している．当初はインターネットを高性能コンピュタあるいは高性能プリンタを利用するように，様々なリソースを遠隔から共有することが主な目的であった．現在は情報の共有，情報の取得といった情報のやり取りが中心となっている．それに伴って，通信形態も変化している．従来の TCP/IP はホスト中心の Host-to-Host の通信形態であり，IP プロトコルは位置情報であるネットワークアドレスを用いてホストアドレスを指定するというロケーション・オリエンテッド*1な通信であった．ところが，現在は情報をユーザに送るというインフォメーション・セントリック*2な通信形態に変わりつつある．

このように TCP/IP の通信形態と現在のインターネットに求められている通信形態との間の差が広がっている．情報の効率的な取得のために P2P*3や CDN*4などの新しいプロトコルが提案された．しかし，これらはロケーション・セントリックな TCP/IP ネットワーク上のプロトコルであるので本質的な解決ではない．本来，情報を取得するという行為に対して，ネットワークアドレスやホストアドレスなどを意識する必要はなく，もし近くにある通信機器が当該コンテンツ (情報)*5を持っておりそこから情報を取得できるなら，それはより効率的であり将来の通信量増大にも対応できると考えられる．そこで, 情報を効率的に取得するために情報指向ネットワーク: Information-Centric-Network (ICN)[3] というプロトコル体系が提案された [4]．

---

*1 Location-oriented: 地理的指向な

*2 Information-Centric: 情報指向な

*3 Peer to Peer: インターネットにおいて一般的に用いられるクライアント・サーバ型モデルでは，データを保持・提供するサーバとそれに対してデータを要求・アクセスするクライアントという 2 つの立場が固定されているのに対して，各ピアに対して対等 にデータの提供及び要求・アクセスを行う自立分散型のネットワークモデル

*4 Content Delivery Network: 頻繁に使われる Web サイトがあると一つのノード（サーバ）だけでは耐えきれないのでいくつかのノードにデータを分散しておき，各ユーザは分散したノードに接続して情報を取得するという方法

*5 参考文献 [3] では情報 (Information) とコンテンツ (Contents) は同様の意味で用いられている．本稿でも同様の意味で用いる．

1.2 情報指向ネットワーク

情報指向ネットワーク (ICN) においてユーザはサーバの IP アドレスではなくコンテンツ名を指定してコンテンツ取得要求を行い，そのコンテンツ要求を受け取った近隣のルータやノードが当該コンテンツを保持していた場合，それらはユーザに対してそのコンテンツを直接転送するプロトコル体系である．ICN において情報を保持している者をパブリッシャ (Publisher)，情報の取得要求を出すものをサブスクライバ (Subscriber) と呼ぶ．また ICN では各コンテンツ (情報) に対してコンテンツ名 (名前) が対応付けられている．

ICN へのアプローチとして様々なが研究がなされているが，現在最も多くの研究者により研究されている Named Data Networking (NDN)[5] 及びその前身である Content Centric Networking (CCN)[6] を代表的な ICN アーキテクチャとして述べる．CCN アーキテクチャはパロアルト研究所[*6]により研究されている ICN の先駆となった本格的なアーキテクチャである．また，US Future Internet Architecture プログラム[*7]によって資金提供された NDN プロジェクトは，CCN アーキテクチャをさらに発展させたものである．

■コンテンツ名 NDN におけるコンテンツ名の命名規則は階層構造になっており，現在のインターネットで流通している識別子である Uniform-Resource-Locator (URL) に似ている．たとえば，コンテンツ名は `/aueb.gr/ai/main.html` となる．ただし，コンテンツ名は必ずしも URL とは一致せず，最初のセクション[*8]は DNS 名または IP アドレスなどの形式である必要もない．つまり NDN では，各セクションについての具体的な規格は定義されていない．またコンテンツ取得要求において，要求されたコンテンツ名のプレフィックスの名前を持つ情報と一致すると見なされる．たとえば，`/aueb.gr/ai/main.html/_v1/_s1` は `/aueb.gr/ai/main.html` という名前のコンテンツと一致する．これは要求されたコンテンツの初版であり，コンテンツを分割したセグメントの最初のデータを表している．このデータを受信したあとに Subscriber は `/aueb.gr/ai/main.html/_v1/_s2` により次のセグメントを要求することや，新たなバージョンを要求することもできる．このようにコンテンツを分割して扱う際にはコンテンツ名にそのメタデータを付与することが可能である．

■名前解決とデータルーティング NDN において，Subscriber はコンテンツを取得する際はコンテンツ取得要求である INTEREST パケットを発行して，Publisher からの DATA パケットの形式で到着するコンテンツを取得する．INTEREST/DATA パケットは，それぞれ要求/転送されるコンテンツのコンテンツ名を持つ．Fig. 1.1 に示すように，すべてのパケットは Content Router (CR) によってホップバイホップ (hop by hop) で転送され，各 CR には 3 つのデータ構造 (Forwarding Information Base (FIB), Pending Interest Table (PIT), Content Store (CS)) がある．FIB は，INTEREST パケットを適切なデータソースに転送するために使用するインターフェイスとコンテンツ名をマッピングする．PIT は，保留中の INTEREST パケットが到着した受信インターフェイス，つまり一致する DATA パケットが転送されてきたときに返送するインターフェースとコンテンツ名をマッピングすることで INTEREST パケットを追跡する．最後に，CS は CR を通過したコンテンツのローカルキャッシュとして機能する．

INTEREST パケットが到着すると，CR はコンテンツ名を抽出し要求されたプレフィックスと一

[*6] Palo Alto Research Center (PARC)：アメリカ合衆国のカリフォルニア州パロアルトにある研究開発企業
[*7] NSF FUTURE INTERNET ARCHITECTURE PROJECT (http://www.nets-fia.net/)
[*8] "/" で区切られた部分をセクションと呼ぶ

致する名前を持つ CS のコンテンツを探す．CS でキャッシュが見つかった場合，すぐに DATA パケットとして受信インターフェイスを介して返送され，INTEREST パケットは破棄される．それ以外の場合，CR はこの INTEREST パケットを転送するインターフェースを決定するために，FIB で最長プレフィックス検索を実行する．FIB でエントリが見つかった場合，CR は PIT に INTEREST パケットの受信インターフェイスとコンテンツ名を記録し，FIB が示す CR に INTEREST パケットを転送します．Fig. 1.1 では，Subscriber は `/aueb.gr/ai/new.htm` という名前の INTEREST パケットを送信する (矢印 1〜3)．PIT にコンテンツ名のエントリが既に含まれている場合，つまりこのコンテンツが既に要求されている場合，CR は受信インターフェイスをこの PIT エントリに追加し，INTEREST パケットを破棄する．

要求されたコンテンツ名に一致するコンテンツが Publisher または CS で見つかった場合，INTEREST パケットは破棄され，コンテンツは DATA パケットとして返送される．この DATA パケットは，PIT で維持されている状態に基づいてホップバイホップ方式で Subscriber に転送される．具体的には，CR は DATA パケットを受信すると，対応するコンテンツを CS に保存し，PIT で最長プレフィックス検索を実行して，DATA パケットに一致するエントリを見つける．PIT エントリに複数のインターフェイスがある場合，DATA パケットが複製され，マルチキャスト配信が実現される．最後に，CR は DATA パケットをこれらのインターフェイスに転送し，PIT からエントリを削除する (矢印 4〜6)．PIT に一致するエントリがない場合，CR は DATA パケットを重複データとして破棄する．NDN では，DATA パケットは INTEREST パケットによって PIT に残された経路に従うため，名前解決とデータルーティングは対称である [7]．

■ICN の実用化に向けて ICN における Contents Router (CR) は未だに研究段階にありソフトウェアとして参照実装[*9]はあるが，ハードウェアとして実装されたものはない．ICN の実用化を行うため，CR のハードウェアによる実装が不可欠である．ハードウェア実装に向けた課題として，TCP/IP における IP アドレスの代わりにコンテンツ名を用いて名前解決とルーティングを行うため CR での処理が複雑であり，各テーブルに必要な記憶容量も増加するという点が挙げられる．この問題を解決するためにコンテンツ名に対してハッシュ化[*10]を行うことによりルーティングに用いる識別子の圧縮などが研究されている [8][9]．しかし，これらは既存のハッシュアルゴリズム[*11]を用いているため，ハードウェアで実装するには複雑過ぎる．また，ICN におけるコンテンツ名のハッシュ化という用途では完全に衝突のないハッシュを用いるより，多少の衝突を許容しながらも高速に計算可能なハッシュアルゴリズムが求められる．この多少の衝突を許容するために各テーブルにおける新たなデータ構造による検索手法も求められる．

1.3 本研究の目的

本研究では，上記課題を解決するために新たなデータ構造による検索手法と高速で軽量なハッシュアルゴリズムを提案，検証することである．コンテンツ名はランダムな文字列ではなくある程度自然言語的な規則があるのでそれを用いることで軽量化を図る．

[*9] Cefore: `https://cefore.net/` NICT により開発された CCNx 準拠の TCP/IP 上で CCN パケットをシミュレートするソフトウェアルータ

[*10] 本稿ではあるデータを規則に則って処理したときに出力されるものをハッシュ，その規則をハッシュアルゴリズム，またこの処理を行うことをハッシュ化と呼ぶ．

[*11] MD5, SHA1 など

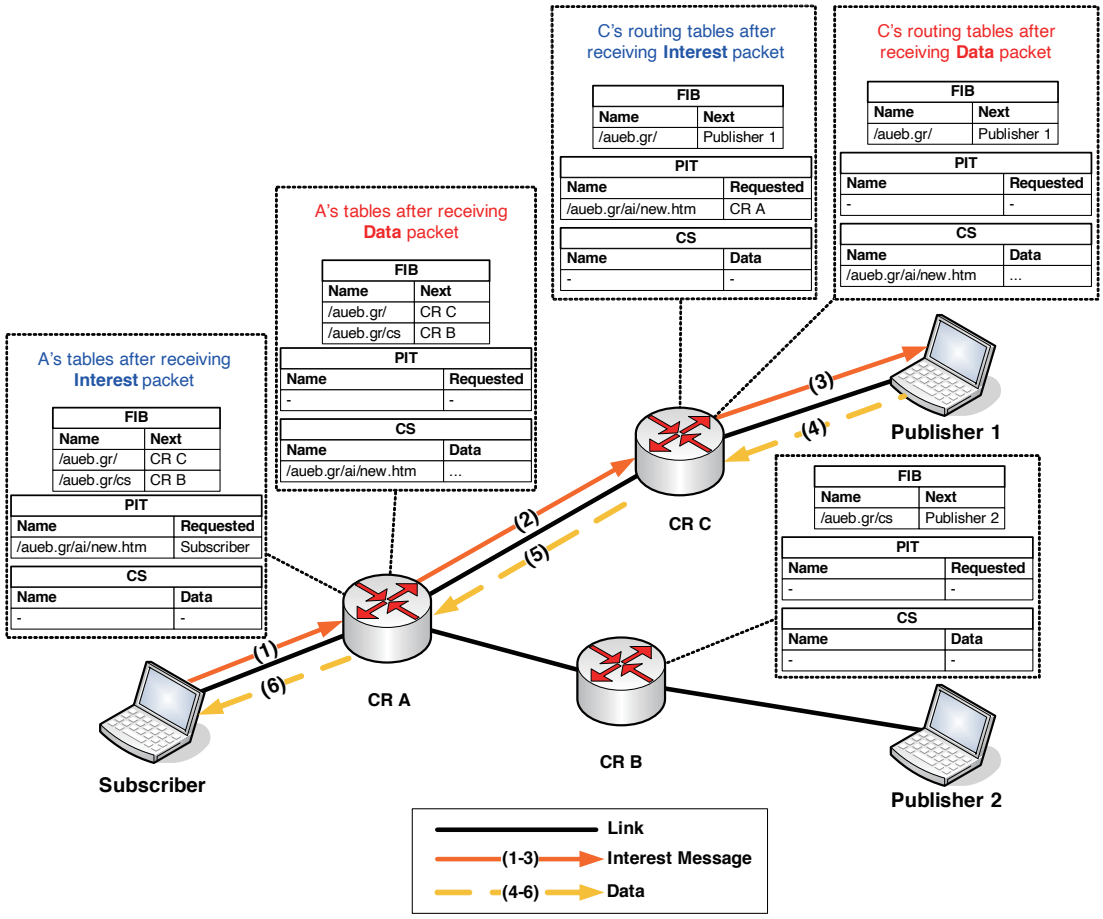

Fig. 1.1 The CCN/NDN architecture. CR stands for Content Router, FIB for Forwarding Information Base, PIT for Pending Interest Table, CS for Content Store (Excerpt from[7]).

第 2 章

解析手法

2.1 URL の構造

NDN のコンテンツ名は URL に似ているので URL の階層構造について調べる．

Uniform-Resource-Locator(URL) は RFC1738[10] により規格化されており，リソースの位置を特定するために用いられる．HTTP URL の構造は

```
<scheme>://<host>:<port>/<path>?<searchpart>
```

である．また，Table 2.1 に各変数の説明を示す．

Table 2.1 HTTP URL variables description[10]

Variables	Description
`<scheme>`	`http`, `https` などのプロトコル体系を表す．
`<host>`	ネットワークホストのドメイン名，または IP アドレス．ドメイン名は”.”によって区切られるドメインラベルの連続．
`<port>`	接続先のポート番号．デフォルトのポート番号 (80,443) 以外のポートを指定するときのみコロンで区切って続ける．
`<path>`	HTTP セレクタでリソースの位置を指定する．
`<searchpart>`	検索パラメータである．これが省略されたときは”?”も省略される．

URL は `host` 部によって階層構造に分けることができる．`host` 部はドメイン名または IP アドレスによって構成されているが，ここでは一般に多く使われているドメイン名のみに注目する．Domain Name System (DNS) は木構造でありドメイン名にはその階層構造が反映されている．前述のドメイン名とは実際には Fully Qualified Domain Name (FQDN)[11] のことであり，Fig. 2.1 のような構造である．FQDN は”.”で区切られた各ラベルの連続であり右端の”.”がルートとなる．ルートから順に各ラベルは Top-Level Domain (TLD), Second-Level Domain (SLD), 3rd-Level Domain (3rdLD), 4th-Level Domain (4thLD), ... のように名前がつけられている．また左端のラベルを Host Name と呼び，それ以外のラベルをまとめて Domain Name と呼ぶ．

Top-Level Domain (TLD) には country code TLD (ccTLD), generic TLD (gTLD), restricted generic TLD (grTLD), sponsored TLD (sTLD) などがありそれぞれ ICANN により管理されている [12]．それぞれの例を Table 2.2 に示す．

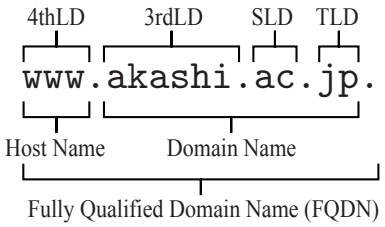

Fig. 2.1 Fully Qualified Domain Name (FQDN) structure. FQDN ends with a period. TLD stands for Top-Level Domain, SLD for Second-Level Domain, 3rdLD for 3rd-Level Domain, 4thLD for 4t-Level Domain.

Table 2.2 Example of each TLDs

ccTLD	gTLD	grTLD	sTLD	eTLD
.jp	.com	.biz	.aero	.co.jp
.uk	.info	.name	.asia	.ac.jp
.cn	.net	.pro	.cat	.akashi.hyogo.jp

■Public Suffix (eTLD) 各ブラウザによるクッキーの適応可能範囲決定のために”1 つのサイト”の単位というものが求められた．それを受けて Public Suffix というものが提案された[*1]．これは effective TLD (eTLD) とも呼ばれ，TLD や co.jp などの実質的に TLD のように機能するドメインを指す．TLD 内のサブドメインの構造は TLD ごとに異なるため，機械的に eTLD を決定することはできない．そのため Public Suffix List (PSL)[13] として一覧表が管理されている．

2.2 ICN におけるコンテンツ名

ICN におけるコンテンツ名を本研究では

icn:/<reTLD>/<Root>/<rHostName>/<Path>

のように定義し，ICN-URL と呼ぶ．reTLD (reverse-eTLD) と rHostName (reverse-HostName) はそれぞれ eTLD と HostName を”.”を区切りとして逆順に配置したものである．すなわち，eTLD が ab.cd.ef なら reTLD は ef.cd.ab となる．Fig. 2.2 に具体例を示す．

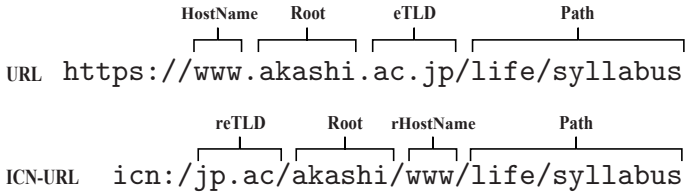

Fig. 2.2 Example of ICN-URL

[*1] 例えば，example.co.jp や a.example.co.jp の Public Suffix は co.jp で，example.co.jp がサイトの単位となる．もし，example.co.jp がドメインが co.jp や jp のクッキーを発行できてしまうと，異なるサイトであるはずの test.co.jp にも干渉できてしまう．これを防ぐためにクッキーの処理では PSL を参照するようになっている．

2.3 性能の評価手順

Fig. 1.1 のテーブルを Fig. 2.3 に再掲する．FIB, PIT, CS の 3 つのテーブルは Name をキーとしており，その分布を評価することでアルゴリズムの性能を評価する．評価のために Fig. 2.4 に示す手順を行った．入手可能な全 URL のリスト (All list) から 10MB 程度のサイズになるようにランダムに抽出したリスト (Sampled list) を作る．そのリストで URL の規格に合わないものを除外したのち，ICN-URL に変換し eTLD の頻度の多い順に並べる．順に ICN-URL からハッシュを計算してハッシュテーブルを作成し，ハッシュテーブルの各 eTLD の先頭アドレスを保持したポインタテーブルを作成する．

FIB

Name	Next
/aueb.gr/	CR C
/aueb.gr/cs	CR B

PIT

Name	Requested
/aueb.gr/ai/new.htm	Subscriber

CS

Name	Data
-	-

Fig. 2.3 Tables of FIB, PIT, CS.

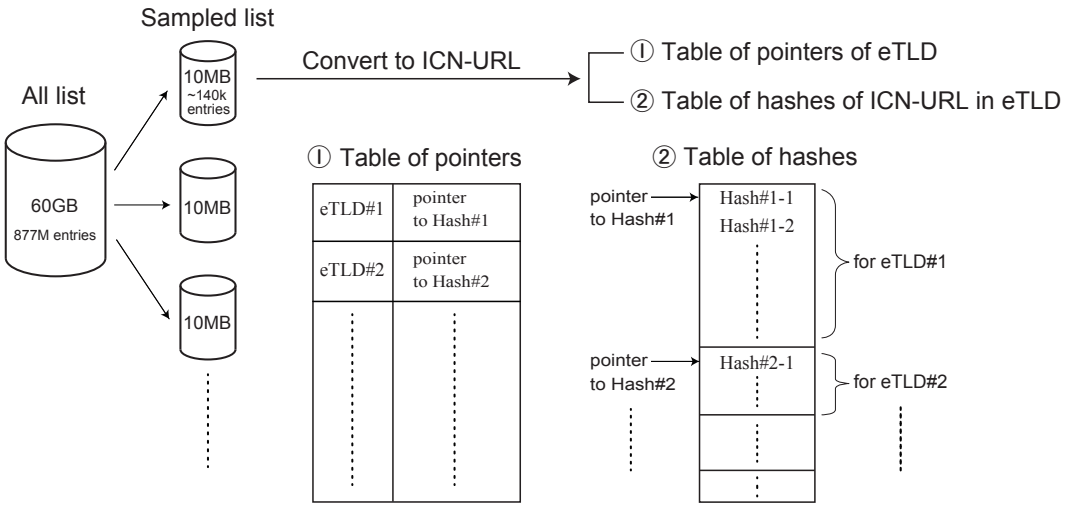

Fig. 2.4 Analysis procedure

2.3.1　解析データ

解析に用いるデータとして，The Content Name Collection[*2]で公開されている情報指向ネットワークのための膨大な URL のデータセットを用いる．そのデータセットの内の一つである `urls.txt` を使う．`urls.txt` は Fig. 2.5 のように改行で区別されている URL のリストになっている．`urls.txt` の概要を Table 2.3 に示す．

```
http://www.google.com
http://images.google.com/imgres
http://www.19lou.com
http://www.sfd.com
http://www.baidu.com
http://www.sina.com.cn
http://www.netvibes.com/
http://www.google.com/search
http://images.google.com/
http://wrestlingabrasil.blogspot.com/
...
...
...
```

Fig. 2.5　Sample 10 in the urls.txt

Table 2.3　Dataset ”urls.txt” overview

Dataset	Number of URLs	unique	File size
`urls.txt`	2,144,314,011	no	121 GB (130,782,049,461 Bytes)

`urls.txt` の前処理として以下の工程を行う．このデータには重複が含まれているので重複を削除し，各 CR での実際的な URL は 10MB ほどであるという仮定のうえ，ランダムな 10MB を抽出し，解析データとした．ここには約 14 万件の URL が含まれる．この解析データ中の各 URL を ICN-URL に変換する．

2.3.2　ハッシュアルゴリズム

ICN-URL からハッシュ値を求めるアルゴリズムを述べる．

まず，Fig 2.6 に示すように ICN-URL を”/”で分割する．それぞれをセクション (section) と呼ぶ．そのセクションの文字数が 3 文字未満の場合はセクションの文字数に応じて 3 文字にする (パディング)．

セクションが 1 文字のとき　ICN の長さとスラッシュの数を掛けたものを uint16 型で付加する

[*2] `http://www.icn-names.net/` にて”The Content Name Collection”というバーゼル大学による情報指向ネットワークのためのデータセットが 2019 年 10 月まで公開されていたが，ドメインの有効期限切れのため現在は全く関係のない中国の会社によりドメインが取得されている．

セクションが 2 文字のとき ICN のスラッシュの数を byte 型として付加する

次に，各セクションから前 3 文字を抜き出して配列 heads とする．同様に後 3 文字を抜き出して配列 tails とする．ただしパティングが含まれているセクションはパディングと元の文字との順序を入れ替える．

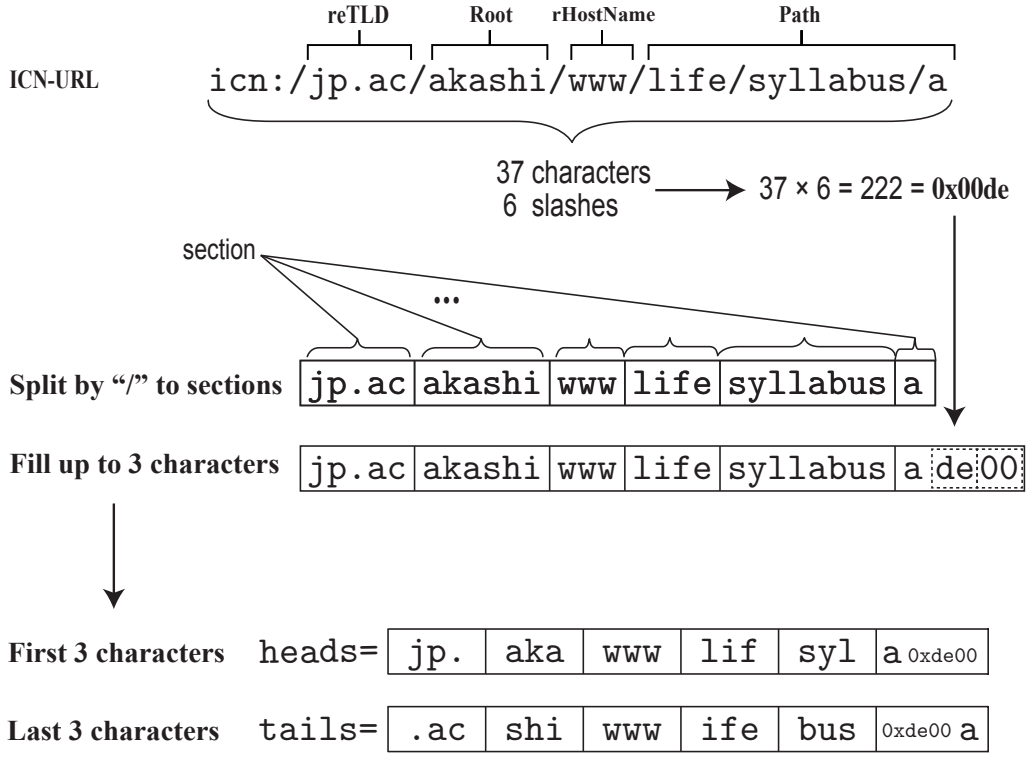

Fig. 2.6 Preparation to make hash.

先程作成した配列 heads と tails を元に 3 種類のハッシュアルゴリズムを考案した．

ハッシュアルゴリズム A

heads の先頭要素 3 バイト，tails の末尾要素 3 バイト，末尾から 3 つ目の要素 3 バイトの各 3 バイト，計 9 バイトをそれぞれのバイトごとに XOR を計算して 3 バイトにする．heads の先頭 1 文字と先程の 3 バイトを連結したものを 4 バイトのハッシュ値とする．

ハッシュアルゴリズム B

heads の先頭要素 3 バイト，tails の末尾要素 3 バイトのそれぞれのバイトごとに XOR を計算し，それを連結して 2 バイトにする．heads の先頭 1 文字と先程の 2 バイトを連結したものを 3 バイトのハッシュ値とする．

ハッシュアルゴリズム C

heads の先頭要素 3 バイト，tails の末尾要素 3 バイト，末尾から 2 つ目の要素 3 バイトのそれぞれのバイトごとに XOR を計算し，それを連結して 3 バイトにする．heads の先頭 1 文字と先程の 3 バイトを連結したものを 4 バイトのハッシュ値とする．

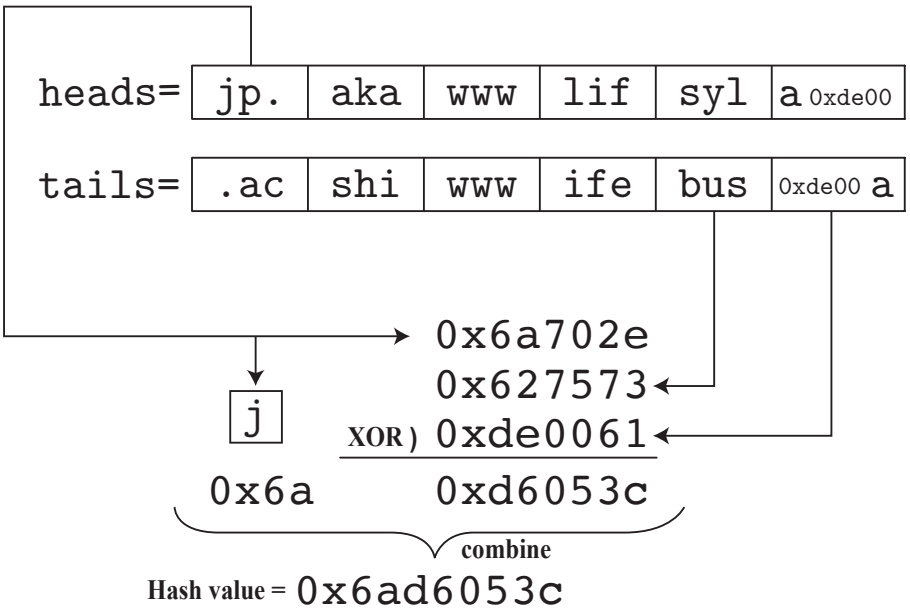

Fig. 2.7　Hash algorithm A

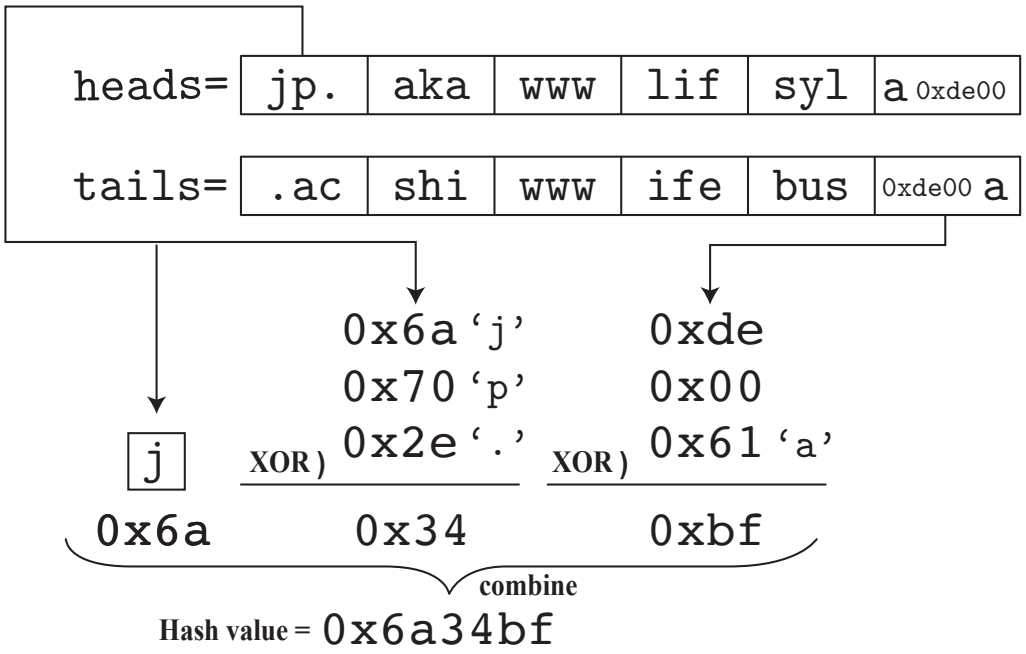

Fig. 2.8　Hash algorithm B

2.3.3　ハッシュテーブル

約 10MB の ICN-URL に変換された解析データを reTLD（1 番目のセクション）の頻度の多い順に並べる．これは，eTLD の頻度の多い順と同じである．各行の ICN-URL に対して上記の 3 種類のハッシュアルゴリズムの内のどれかにより順にハッシュ値を求め，ハッシュテーブルを作成する．このハッシュテーブルには頻度順の同じ eTLD に対応するハッシュ値が連続して並んでいる．すなわ

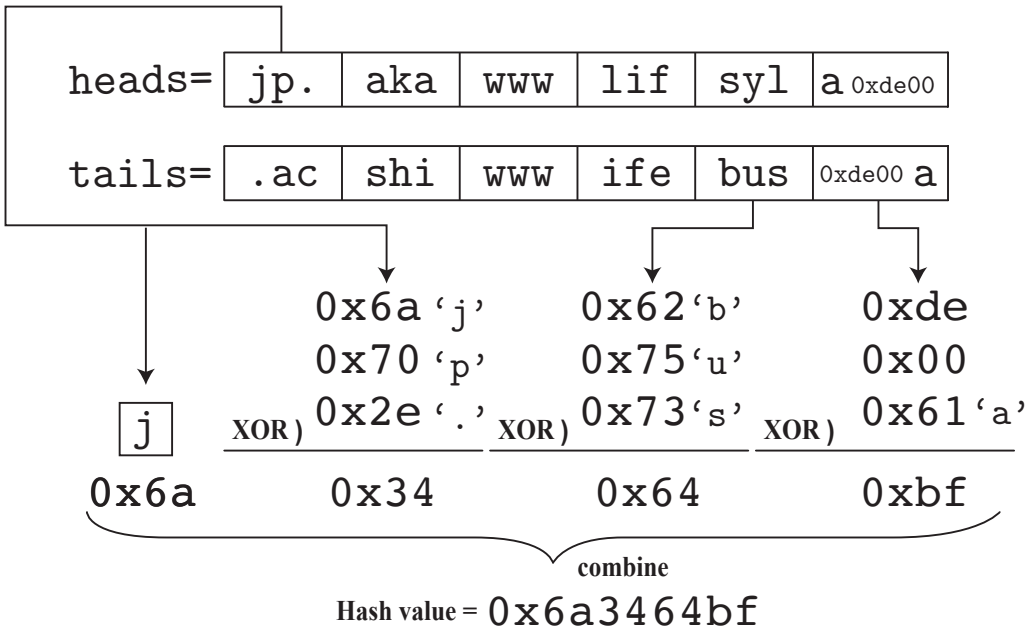

Fig. 2.9 Hash algorithm C

ち，eTLD ごとにグループ分けされたハッシュテーブルが得られる (Fig 2.4 の①)．新たな eTLD のグループの出現する位置のアドレスをポインタと呼び，それと eTLD との対応関係をポインタテーブルと呼ぶ (Fig 2.4 の②)．

2.4 プログラム

ハッシュアルゴリズムの検証・URL の構造解析を行うためにシミュレーションプログラムを Golang で作成した．

第 3 章

衝突数の検証結果

3.0.1　ハッシュアルゴリズム

3.1　URL の分類手法を利用するとき

3.2　ハッシュと URL の分類手法を併用したとき

参考文献

[1] David D. Clark et al. Barry M. Leiner, Vinton G. Cerf. Brief history of the internet. Internet Society, 1997.

[2] Cisco. Cisco visual networking index: Forecast and trends, 2017 - 2022. Cisco, 2019.

[3] Van Jacobson, Diana K. Smetters, James D. Thornton, Michael F. Plass, Nicholas H. Briggs, and Rebecca L. Braynard. Networking named content. In *Proceedings of the 5th International Conference on Emerging Networking Experiments and Technologies*, CoNEXT '09, pp. 1–12, New York, NY, USA, 2009. Association for Computing Machinery.

[4] 朝枝仁, 松園和久. 情報指向ネットワーク技術におけるプロトタイプ実装と評価手法. コンピュータ ソフトウェア, Vol. 33, No. 3, pp. 3_3–3_15, 2016.

[5] NSF Named Data Networking project. [Online]. Available: `http://www.named-data.net/`.

[6] Content Centric Networking project. [Online]. Available: `http://www.ccnx.org/`.

[7] G. Xylomenos, C. N. Ververidis, V. A. Siris, N. Fotiou, C. Tsilopoulos, X. Vasilakos, K. V. Katsaros, and G. C. Polyzos. A survey of information-centric networking research. *IEEE Communications Surveys Tutorials*, Vol. 16, No. 2, pp. 1024–1049, Second 2014.

[8] Lorenzo Saino, Ioannis Psaras, and George Pavlou. Hash-routing schemes for information centric networking. In *Proceedings of the 3rd ACM SIGCOMM workshop on Information-centric networking*, pp. 27–32, 2013.

[9] 顕士小松, 卓也朝香. コンテンツ指向ネットワークにおけるブルームフィルタを用いた経路情報管理方式 (ネットワークシステム). 電子情報通信学会技術研究報告 = IEICE technical report : 信学技報, Vol. 113, No. 35, pp. 13–18, may 2013.

[10] M. McCahill Xerox Corporation. Uniform resource locators (url), dec 1994. [Online]. Available: `https://tools.ietf.org/html/rfc1738`.

[11] Xylogics Xylogics. Internet users' glossary, aug 1996. [Online]. Available: `https://tools.ietf.org/html/rfc1983`.

[12] Internet Assigned Numbers Authority. Root zone database, feb 2020. [Online]. Available: `https://www.iana.org/domains/root/db`.

[13] Mozilla Foundation. Public suffix list, feb 2020. [Online]. Available: `https://publicsuffix.org/`.