【様式4】

テーマ名：P2Pアーカイブ：P2P型Webアーカイブシステムの構築
申請者名：札場 寛之、森下 睦

【提案テーマ詳細説明】

1. なにを作るか

目的

本プロジェクトの目的は、従来のWebアーカイブシステムをP2Pネットワーク上で実現することである。従来のWebアーカイブシステムは、大量のWebページをクローリングし、多くのユーザにコンテンツを配信する必要がある。これを実現するためには、大容量のストレージ、潤沢なネットワーク帯域及び計算資源が必要であるが、既存のサービスは資金面を募金に依存しており、これらが将来的に維持し続けられる保証が無い。本プロジェクトでは、Webアーカイブシステムをユーザをノードとしたp2pネットワーク上で実現することで、これらの問題を解決する。ユーザそれぞれが少しずつ負担を共有し、ランニングコストのかからない存続可能なサービスを目指す。またこのシステムをP2Pネットワークを用いたコンテンツデリバリネットワーク（CDN）として使うことも意図している。これにより、Webサーバに負荷をかけずにWeb上の最新のコンテンツをより高速にダウンロードすることが可能になる。

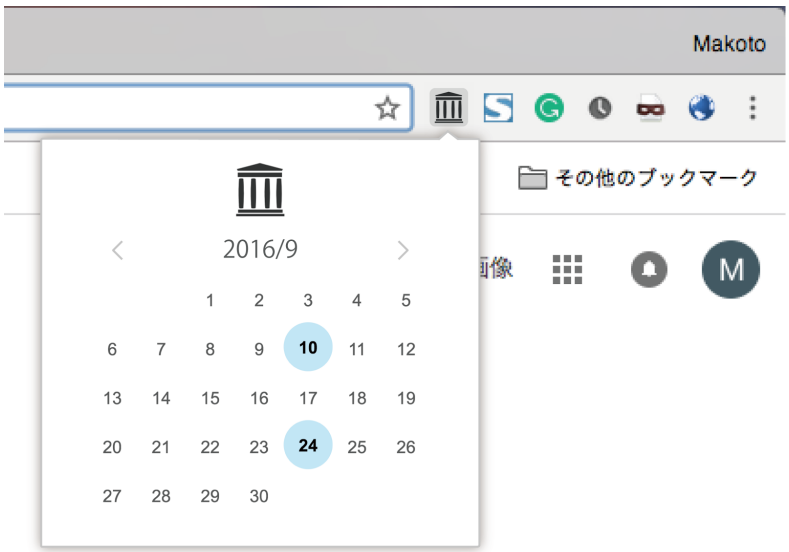

図1　ブラウザプラグインとして動作している本プログラムのイメージ図。ブラウザプラグインとして公開することで、導入の敷居が下がりより多くのユーザが見込める。詳細は2節で述べる。

図2　CDNとして本プログラムを利用しているイメージ図。P2Pネットワークを利用することで、通常のWebサーバからダウンロードする場合と比較してより高速にダウンロードが可能な場合がある。

背景

Webアーカイブ

　Webアーカイブとは、Webサイトをクローリングし、収集したコンテンツをストレージへ保存、これを公開するサービスである。代表的なサービスにThe Internet Archive (http://archive.org/web) などが挙げられる。Webアーカイブの目的は、インターネット上の全てのデータを保存し、全世界の人々が将来に渡って利用できるインターネット上の図書館を作ることである。Webアーカイブを参照することで、削除されたり改変されたWebコンテンツを以前の状態で閲覧することが可能になる。

　しかし、既存のWebアーカイブはクライアント-サーバモデルを採用しているため、以下のような問題がある。

- **可用性の問題**
  ディスク障害等に対しての可用性を確保するためにはバックアップディスクなどを用意する必要があるが、これにはコスト面での問題が残る。
  また、現状The Internet Archiveは募金に頼って運営しており、今後このサービスが永続できる保証が無い。

- **増え続ける容量の問題**
  2012年の時点で、The Internet Archiveに保存されているデータは10PBである。
  資金面を募金に依存したシステムでは、インターネットの拡大に従い今後増加していくコンテンツを網羅するのは困難である。

- **トラフィックの増大・集中**
  The Internet Archiveは2012年時点で、サービスを維持するために20Gbit/s以上の多くのネットワーク帯域を必要としている。
  インターネットの利用拡大に伴い、動画等の大容量コンテンツがインターネット上に幅広く浸透してきているが、これらをWebアーカイブ上に保存、配信するにはコスト面での問題が存在する。
  また、Webアーカイブ上の特定のコンテンツにアクセスが集中する場合があるがこれに対して特別な対策を取っていない。

- **計算資源の問題**
  更新頻度の高いWebサイトだとクロールとクロールの間にコンテンツが変わりデータが失われる可能性がある。
  しかしクローラの更新頻度を上げるためには、より多くの計算資源を必要とする。

CDN

さらに、クライアントとサーバ間のデータ通信は、その物理的な距離と通信経路の状態に大きく依存する。故に、遠く離れたサーバとの通信では、大容量コンテンツをダウンロードする際に長い時間がかかる場合がある。

通常、この問題はCDNを用いると解決できる。CDNはコンテンツを複数のサーバへ置き、リクエストしたユーザにとって最適なサーバからコンテンツを配信することで配信を高速化する。しかし、CDNはランニングコストが高く、維持が難しいという問題がある。

BitTorrent

BitTorrentは大容量のコンテンツを高速に共有するためのプロトコルである。しかしコンテンツを共有するためにはトラッカーサーバの設定やtorrentファイル作成など、アップロード者の操作が必要となり、ダウンロード時もプロトコルをサポートした特別なクライアントを必要とする。

ゆえに、有益な性質を持ちつつもWebコンテンツの負荷分散などへの使用は限定的である。

目標と提案内容

我々はP2Pネットワーク上にWebアーカイブを構築することを目標とする。このシステムでは、ネットワークに参加する各ユーザがWebページを訪問すると同時にクローリングを行い、コンテンツをP2Pネットワーク上で共有することで、前述した既存のWebアーカイブの問題を解決する。

また、Web上のデータを自動的にP2Pネットワーク上で共有することで、特定のWebサーバへ大きな負担をかけることなく、Webコンテンツの高速なダウンロードを実現する。これにより、CDNを実現することも可能である。

本プロジェクトのP2Pネットワークは、以下に説明するPUTとGETの2つのリクエストをサポートするノードによって構築される。各ユーザからP2Pネットワークへコンテンツを追加するリクエストをPUTと呼び、P2Pネットワーク上からコンテンツを取得するリクエストをGETリクエストと呼ぶ。

図3、4にそれぞれのリクエストのイメージ図を示す。以下にそれぞれの具体的な動作を示す。

PUTリクエスト

1. ユーザがWebサーバにアクセス
2. Webサーバがコンテンツを返す
3. クライアントはP2Pネットワーク上のコンテンツが最新か問い合わせる
4. P2Pネットワーク上にあるコンテンツが古い場合、最新のものをアップロードする
5. P2Pネットワークは担当するノードへコンテンツを保存する

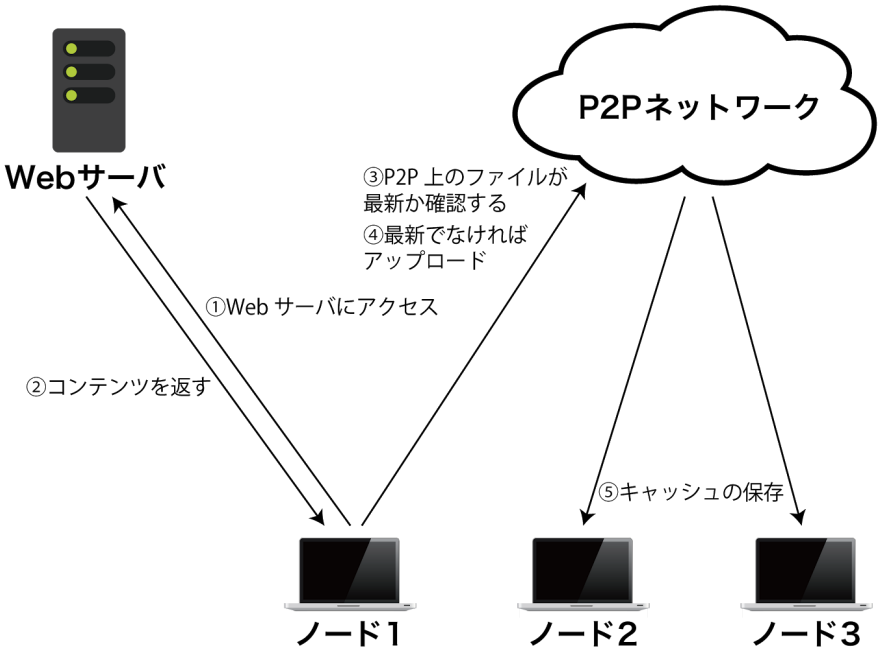

図3　PUTリクエストの動作イメージ図

GETリクエスト

1. ユーザがWebコンテンツにアクセス
2. Webコンテンツがリンク切れ、あるいはユーザが過去の状態を閲覧したい
3. URLをキーにP2Pネットワークにコンテンツを問い合わせ
4. P2Pネットワーク上のノードがコンテンツを返す

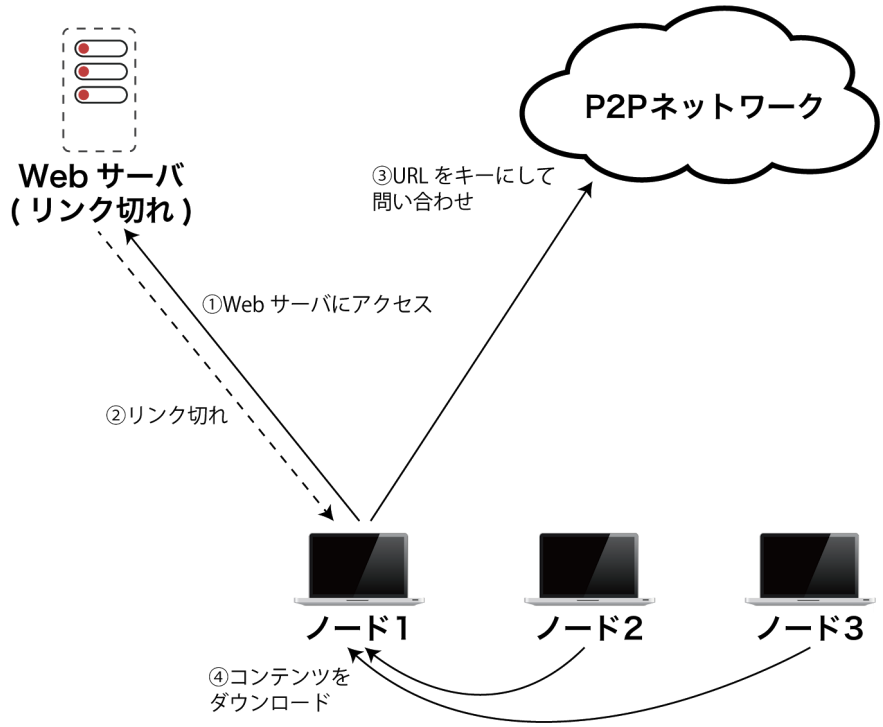

図4　GETリクエストの動作イメージ図

本プロジェクトで構築するP2Pネットワーク

本プロジェクトで構築するP2Pネットワークのイメージ図を図5に示す。本プロジェクトでは前述の2つのリクエストを実現するため、分散ハッシュテーブル (DHT) 及びBitTorrentネットワークからなる2つの異なるオーバーレイ・ネットワークを用いる。

DHTの代表的なアルゴリズムとしては、ChordやKademliaなどがある。提案手法では、これらを用いてコンテンツを持つノードがどこにあるかを検索する。各ノードはそれぞれBitTorrentでのトラッカーのような役割を担う。このDHTネットワークは言い換えればコンテンツを保持するノードのアドレス群を、ハッシュ値化したURLから引く辞書である。そうして得られたノードの情報を元に、BitTorrentプロトコルを用いて実際にコンテンツのダウンロードを行う。より多くのノードが同一のコンテンツを保持するように設定すれば、本Webアーカイブ全体の可用性が向上する。

また、本P2Pネットワークを利用することで、Webサーバ上にファイルが存在する場合でも、P2Pネットワーク上に同一のファイルが存在すれば、ここからファイルをダウンロードすることでBitTorrentの利点を活かし高速にコンテンツをダウンロードすることが可能になる。このように、本ネットワークはWebアーカイブだけでなくCDNとしても応用可能である。

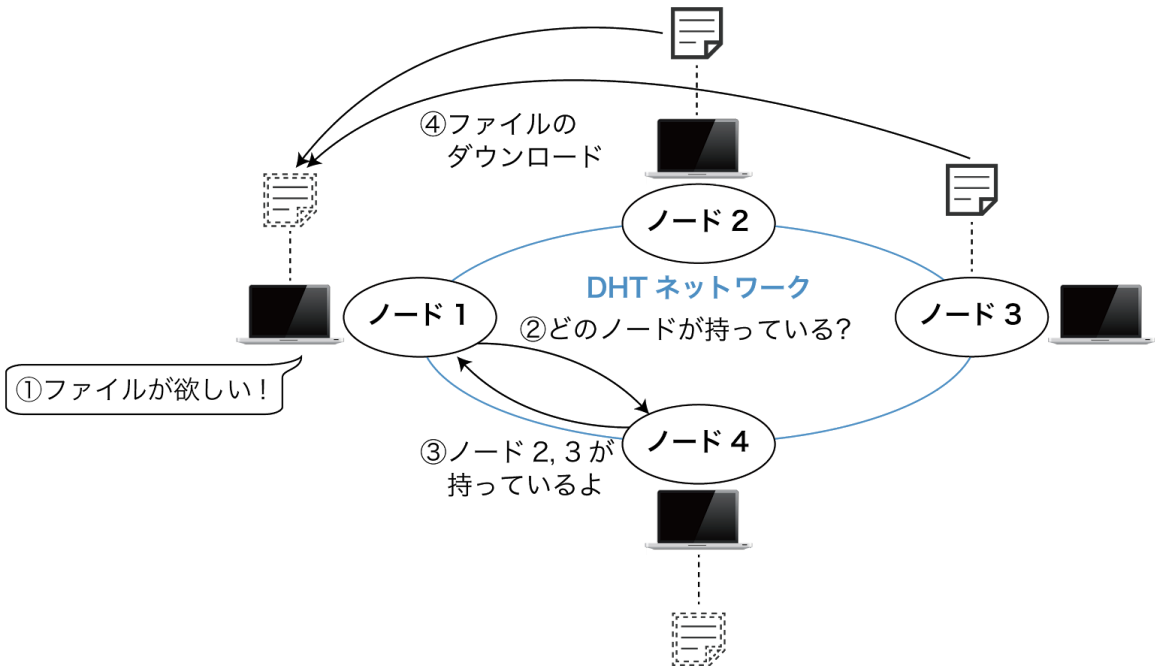

図5　本プロジェクトで構築するP2Pネットワークのイメージ図

2. どんな出し方を考えているか

本システムは、ユーザ数が増加するに従いコンテンツ配信は高速になり、また一人あたりの負担は減ると考えられるため、極力多くのユーザに本システムを使ってもらえることが重要である。

そこで、我々はピア実装をGoogle ChromeなどのWebブラウザのプラグインとして公開することを目指す。Webブラウザのプラグインは既存のBitTorrentクライアントなどに比べユーザが容易に導入できユーザ数の拡大が狙いやすい。また、図1、2のような使いやすく

わかりやすいUIを心がけることで、一般のユーザが利用しやすい環境を整える。またGETリクエストを行わない常駐型のアプリケーションも提供することで、任意のサーバにボランティアへの参加を促す。

また、幅広いユーザにリーチできるようプロジェクトの広報活動を重視する。これには未踏の持つブランド力を最大限に利用していきたい。ユーザが実際にプラグインやアプリケーションを使ってその利便性を体感できるように、デモ用のWebページやコンテンツを用意する。これらをプロジェクト早期から実現するために、ソフトウェア開発手法としてプロトタイピングモデルを採用し、極力早い段階でベータ版をリリースする。

ソースコードはGitHubなどで公開し、誰でも閲覧・改変できるようにする。また開発に用いた技術に関する知見を日本語のドキュメントとして残すことで、日本のIT関連産業の発展に貢献する。

3. 斬新さの主張、期待される効果など

本プロジェクトの最も斬新な点は、P2Pを用いてWebアーカイブを各ユーザの負担によって実現する点である。既存のWebコンテンツの内容を1つの大きなP2Pネットワーク上で共有するシステムは我々の知る限り存在せず、提案手法では既存のWebアーカイブに存在する多くの問題を解決できる。

- **可用性の問題**
  P2Pネットワーク上で複数のノードに同一のコンテンツを保存することで、一つのノードを失った場合でもサービスを続けられる。また、The Internet Archiveのように一つの運営母体に依存しないため、ユーザのみでサービスの継続が可能である。

- **増え続ける容量の問題**
  各ユーザが公平に余剰ストレージを提供することで、1つのマシンに多くのストレージを用意する必要がなくなる。

- **トラフィックの増大・集中**
  P2Pネットワークを利用するため、一つのノードにアクセスが集中する事態を避けられる。また、大容量のファイルであっても複数のノードが協力してピースを保持、提供することで、各ユーザあたりの負担が大きくならないようにする。

- **計算資源の問題**
  各ユーザ自身がクローラとなるため、運営者が大規模なクローラや計算資源を用意する必要がない。ユーザがアクセスしたコンテンツをP2Pネットワーク上に保存するため、より多くの需要があるコンテンツが優先的に短い間隔で保存される。

また、BitTorrentの特徴として多くのノードに保持されているコンテンツはより高速にダウンロードが可能であるという点がある。本プロジェクトで構築するシステムではこの特徴を利用して、Web上の多くのファイルを本来のWebサーバからではなく、P2Pネットワークからもダウンロードできるようにすることで、より高速な通信を実現できる点も斬新である。

将来的には、この技術を応用することで従来Webサーバに置かれていた情報を全てP2Pネットワークに移動させることも可能になると思われる。これが実現すればWebサービスが現在のクライアント-サーバモデルからP2Pネットワークモデルへと全面的に移行できる

ようになり、Webサービスにサーバを一切使用しない次世代のネットワークが構築可能となる。

CacheP2P.jsとの違い

Web上でBitTorrentプロトコルの通信を実現するWebTorrentのようなライブラリや、それを元にしたCacheP2P.jsがある。CacheP2P.jsは、Webページ管理者がWebページ中に埋め込むことで、そのページを閲覧しているユーザ同士がP2Pネットワークを構築し、CDNを実現するライブラリである。これを用いればサーバへの負荷を軽減することができ、また高速にコンテンツを配信することができる。しかし、CacheP2P.jsはユーザがそのページを閲覧している時のみ動作するため、可用性やストレージ容量などの問題を解決するわけではない。

Freenetとの違い

Freenetは匿名性やセキュリティを重視したP2Pネットワークである。FreenetはP2Pネットワーク上でWebコンテンツを配信するという点で本プロジェクトと類似している部分がある。しかし、Freenetはコミュニケーションを匿名かつ安全に行うことに重きを置いており、利便性や速度面の効率性は重視していない。また、このネットワークを利用するためには特別なブラウザやその他フロントエンドの導入が必要となりユーザにとって敷居が高い。

4. 具体的な進め方と予算

(1) 開発を行う場所
　　自宅
(2) 使用する計算機環境
　　Linux、MacOS、AWS
(3) 使用する言語、ツール
　　言語: JavaScript、Python3、C++
　　ライブラリ: WebTorrent
(4) 各クリエータの作業の分担
　　札場：プラグイン、アプリケーション側全般
　　森下：ネットワーク基盤設計および実装
(5) ソフトウェア開発に使う手法
　　プロトタイピング
　　GitHub上でのバージョン管理
(6) 開発線表
　　① 6月中旬～7月中旬
　　クローラのプロトタイプ実装
　　最新のコンテンツをローカルに保持
　　α版の公開
　　② 7月中旬～9月初旬
　　DHTとBitTorrentクライアントの実装
　　コンテンツを分散して保持
　　③ 9月初旬～10月初旬
　　ブラウザプラグインの実装

④　10月初旬～11月中旬
バグ修正
広報活動
β版の公開

⑤　11月下旬～2月下旬
β版のフィードバックを元に改善

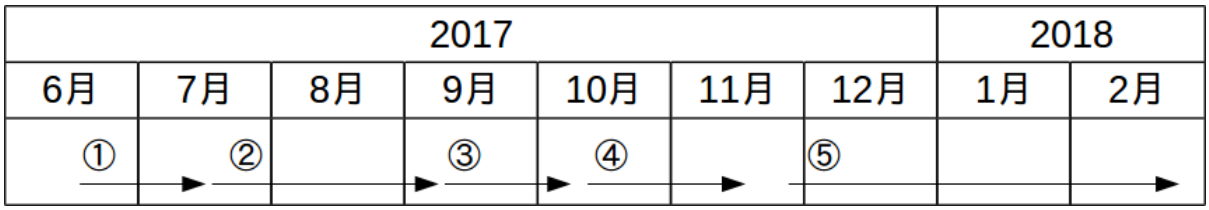

図6　開発線表

(7) 開発に関わる時間帯と時間数
月曜日～金曜日：1日2時間、終業後の20時から22時まで
土曜日～日曜日：1日5時間、13時から18時まで
1人あたり週20時間の開発を予定している。

(8) 予算内訳をまとめた表

名前	時給	時間	必要予算
札場 寛之	1,600	720	1,152,000
森下 睦	1,600	720	1,152,000
合計		1,440	2,304,000

5. 提案者たちの腕前を証明できるもの

札場 寛之

プログラミング言語
Clojure、Python、JavaScript

専門
機械翻訳、自然言語処理、自動プログラミング
ソフトウェア工学トップカンファレンス採択など
Pythonから日本語への統計的変換（https://github.com/delihiros/pseudogen）

株式会社テンクーにてエンジニアのアルバイト
ネットワーククラスタリングなどの実装（https://github.com/xcoo/gugus）

その他
TOEIC 925点を取得し、英語が問題なく使用できる。
プログラミング言語Futhonのフルスクラッチ実装
（https://github.com/delihiros/futhon）
マンガを用いたチャットサービスChatomicの開発（http://parse.jp）
Webページ（http://delihiros.jp）

森下 睦

プログラミング言語

Python、C++

専門

機械翻訳、自然言語処理
これまでに、論文誌1本、国際会議2本、国内会議5本の発表を行った。
深層学習を用いた文間類似度判定をフルスクラッチで構築
（https://github.com/MorinoseiMorizo/sentence_similarity）
深層学習を用いたオープンソース機械翻訳ツールキット「NMTKit」へのコントリビューション (https://github.com/odashi/nmtkit)

資格

情報セキュリティスペシャリスト取得
ネットワークスペシャリスト取得
データベーススペシャリスト取得

その他

TOEIC 970点を取得し、英語が問題なく使用できる。
マンガを用いたチャットサービスChatomicの開発（http://parse.jp）
Webページ（http://www.otofu.org）

6. プロジェクト遂行にあたっての特記事項

札場 寛之

2017年4月よりA社に就職予定。採択が決定した場合、本事業による支援措置を受けること及び開発成果がクリエータ個人に帰属することについて、所属組織から承諾を受ける。

森下 睦

2017年4月よりN社に就職予定。採択が決定した場合、本事業による支援措置を受けること及び開発成果がクリエータ個人に帰属することについて、所属組織から承諾を受ける。

7. ソフトウェア作成以外の勉強、特技、生活、趣味など

札場 寛之

性格：一度エンジンがかかると食事も取らずに気が済むまでのめり込む。興味の範囲が広く自然言語処理、ネットワーク、画像処理等幅広い分野の論文を読み、気になったものは実装する。
趣味：カメラ、ボルダリング、散歩

森下 睦

性格：何事も堅実にこなすタイプ。リーダーとして周囲の人を管理してプロジェクトを進めていくのが得意。札場はエンジンがかかると早いもののエンジンのかかりが遅い時があるので、森下がスケジュールを見て点火することが多い。興味の幅広く、情報処理技術者試験の高度区分に複数合格している。
趣味：ボルダリング、散歩、サイクリング

8. 将来のソフトウェア技術について思うこと・期すること

本節では、今後のソフトウェア技術の発展について重要だと思われる点について我々の意見を述べる。

我々が最も重要だと考える点は、ソフトウェア技術と他分野の融合である。今日、様々な分野でソフトウェア技術が応用され、これまでに無かったサービスや新たな仕組みの開発などが進んでいる。この一例として、ソフトウェア技術と出版業界が融合し、電子書籍が生まれたことが挙げられる。このように一部の分野ではソフトウェア技術を活用することで、これまでの常識を大きく覆す変化が起きている。

しかし、まだソフトウェア技術が浸透していない分野も数多くある。農業などがその一例である。今後このような分野に対してもソフトウェア技術を応用し、より良い社会を実現するためにも、ソフトウェアエンジニアと他分野のつながりをより強くし、他分野へソフトウェア技術が応用しやすくなる場作りが大切だと考える。

また、次世代を担う若者の教育も重要だと考えている。近年のソフトウェア技術や情報技術の発展は目覚ましいものがあり、社会全体での重要性は日に日に増している。これに伴いソフトウェア技術に秀でた人材をより多く育成する必要性が生まれてきた。近年では、それを反映するように小学校でのプログラミング教育の必修化が決定したが、これだけに留まらず、中高大学生へのプログラミング教育の充実や、ソフトウェア技術を学びたい学生が自由に学べる環境の整備、社会的な活動支援の拡充などが重要だと思われる。

我々は、今後ソフトウェア技術が社会全体に貢献し、持続的な発展を可能にするために、「ソフトウェア技術と他分野の融合」及び「次世代を担う若者の教育」の2点が重要であると考えている。そのためにも、この未踏プロジェクトを通して得られた成果およびその過程で得られた技術的知見を社会全体に広く公開し、将来のソフトウェア技術の発展に寄与したいと考えている。

参考文献

[1] Internet Archive “Wayback Machine”
http://warp.da.ndl.go.jp/contents/reccommend/world_wa/world_wa02.html
[2] Want to help build a distributed web?
https://blog.archive.org/2012/02/15/want-to-help-build-a-distributed-web/
[3] Chord: A scalable peer-to-peer lookup service for internet applications,
Ion Stoica+, 2001, SIGCOMM '01
[4] Kademlia: A Peer-to-Peer Information System Based on the XOR Metric,
Petar Maymounkov+, 2002, IPTPS '01