第４部

特集4 Bigdata実証実験プラットフォーム

砂原 秀樹、藤枝 俊輔、江原 千尋、高野 祐輝、増田 和紀、石田 慎一、西 宏章

第１章　はじめに

Bigdataに関する研究は各所で行われているが、その中でも比較的早い時期にBigdataの研究に着手したのはWIDE Projectではないかと考える。1997年からはじめたインターネット自動車プロジェクトの活動では、自動車が生成する情報を収集し解析することでさまざまな情報が生成できることを示してきた[47]。またこうした中で、情報の表現形式[48]、あるいは個人情報の保護基本原則[49]といったことを国際標準としてきた。また、2005年からはLive E!プロジェクトにおいて環境情報を収集し、ゲリラ豪雨の検出等に活用することも行ってきている[50]。

一方で、インターネット上を流通する情報は多種多様となり、これらをどのように活用するかについて考える必要が出てきた。そこで、WIDE ProjectにおいてもBigdataを取り扱うための研究プラットフォームを構築することとした。具体的には、合宿を実験場とし参加者の行動履歴等収集できる情報を収集し解析するという取組を行うこととした。

具体的な研究テーマは以下の3つとなる。

a. 情報の収集技術の確立
b. 情報の分析活用技術の確立及び応用の開拓
c. 情報収集にあたって社会受容性の検討

特に、3つ目の項目については社会で広く受け入れられ情報を活用するために不可欠な項目であり、どのようにパーソナル情報の所有者に説明するか等について検討を行う。一方で、こうした問題を解決しなければ一般環境での情報収集は難しいため、WIDE Project内での情報収集・利用に限定することで、これらの検討を並行して行うことができるようにした。

なお、これらの準備は2013年3月の合宿以降進められ5月研究会を経て、9月の合宿において一回目の実験を実施した。なお、準備の関係からまず情報の収集技術を中心に実験を行った。具体的には、WiFiロケーション技術及び音波ロケーション技術による位置情報収集実験、パケット解析によるDeep Packet Inspection及びL7コンテンツ解析を実施している。詳細を次章以降に示す。

第２章　2013年WIDE秋合宿における無線LANによる位置情報取得実験報告

2.1　はじめに

2.1.1　実験概要

本実験では、合宿地において無線LAN通信から端末位置を推定する運用実験を行った。近年、屋内における位置情報や行動履歴を把握するために、専用機器不要で無線LANの電波検出を利用するリアルタイム位置情報システム(Real Time Location System、以下RTLS)の開発が進み、病院や空港など人の動きを把握する需要が高い場所で利用が始まっている。RTLS技術の動向を把握し、今後研究利用するために、市販されているRTLSをWIDE合宿会場で運用した。合宿地のプレナリホール、ワークショップ会場、会議室、それらを結ぶ廊下を対象に、合宿参加者が持つ無線LAN端末の推定位置データを収集した。収集したデータは、生データとしてcamp-netから合宿参加者に提供し、“無線LANいろいろワークショップ”等において参加者の自由な解析・利用ができるようにした。

2.1.2　実験の目的

本実験の目的は、これまで病院や空港といった特殊な需

要に応じて利用されてきた屋内位置情報を、汎用の無線インフラから簡単に取得した場合に、どのような環境が必要で、どの程度の精度が期待できるかを把握し、今後の利用検討に発展させることである。そのため、今回は以下の情報把握を目的としている。

・RTLSの技術要素、システム、機能の把握
・RTLSを動作させる無線LANインフラの構築条件の調査
・無線LANによるRTLSから得られるデータ精度

2.2 実験環境

2.2.1 RTLS

本実験で用いたRTLSでは、端末の送信信号を複数のAP（Access Point）で受信し、各APにおける受信信号強度（Received Signal Strength、以下RSS）から端末位置を推定する。RSS方式は遮蔽物による減衰、雑音、マルチパスなどの影響を受けやすいが、位置推定方式の中では最も簡易に実現でき、検出対象ノードには一般の無線LAN端末を利用できる。今回利用したシスコ社のRTLSでは、無線LAN端末が送信するプローブリクエストを位置検出の対象としている。図2.1のように、APは端末からの信号を検出すると、そのRSSをWLC（Wireless LAN Controller）に送り、WLCは集約したRSS情報をMSE（Mobility Service Engine）に渡す。MSEは複数のAPが報告したRSSから端末の相対位置を計算する。RSS方式は原理的には3点測位であり最低3台のAPで信号を検出する必要がある（図2.1では4つのAPが端末からの信号を受信しており、受信AP数が多いほうが精度が向上する）。RSS方式では、屋内における遮蔽物や反射がRSSに影響するため、検知対象とするエリアの物理情報を位置計算に含めることが望ましいが、時間的な制約と、詳細な物理情報（どの壁や柱が何dB減衰するか）をくまなく取得することは困難であり、また不完全な物理情報がデータを歪ませることを避けるため、今回はエリアの物理情報は含めずに実験を行った。

2.2.2 使用機材

本実験では、ユーザが会場に持ち込んだ一般の無線LAN端末（ノートPC、スマートフォン、タブレット）を対象に位置情報を取得した。位置情報取得システムと無線LANインフラには以下の機材を利用した。

・Mobility Service Engine(MSE)
  - RSSから位置計算を行い、システム管理者や関連機器にその情報を提供する。
  - バージョン7.5.102.0、VMware ESXi上の仮想マシンとして動作
・Prime Infrastructure(PI)
  - ネットワーク統合管理システムであり、システム全般（AP、WLC、MSE）の設定および制御と、システム管理者による情報閲覧に利用する。

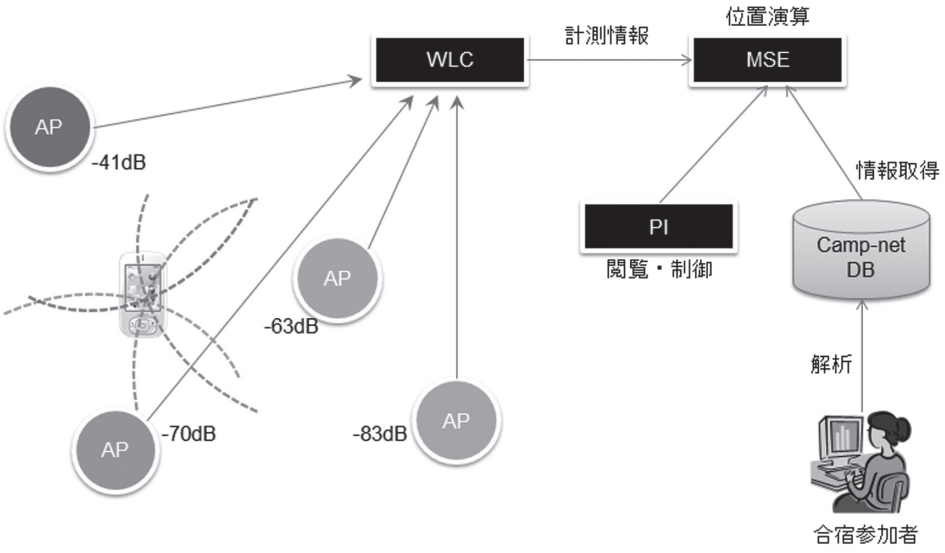


図2.1 実験で用いたシスコ社のRTLSシステム

- バージョン1.4.0.45、VMware ESXi上の仮想マシンとして動作

・ワイヤレスLANコントローラ(Wireless LAN Controller, 以下WLC)

- ファームウェア 7.4.100.0、WLC5508

・無線LAN基地局(802.11a/g/n)

- Aironet 2600 4台
- Aironet 3500 16台

2.3 位置取得に適した電波環境の構築

2.3.1 APの配置

本実験では、合宿参加者に生活用の無線LANインターネットを提供しつつ、端末の位置情報を取得した。無線LANは802.11a/g/nで提供し、以下の電波環境を構築した。APが端末の信号を受信するためには、基地局自身がそのチャネル上で監視状態である必要がある。通信サービスを提供するAP(以下通信AP)は、自分がサービスしているチャネル以外を一定時間毎に周回して監視可能であるが、監視する時間が非常に短いため効果的に端末を検出できない。通信APはデフォルト設定において16秒の通信サービス時間の後に50msの他チャネル監視時間を持ち、通信サービス・他チャネルAを監視・通信サービス・他チャネルBを監視のように他チャネルを順に周回し監視する。デフォルト設定の監視間隔(約180秒)で2.4GHz帯の11チャネル全体を監視することになる。通信APでは他チャネルを検出できる時間が非常に短いため、監視専用のAP(以下モニタAP)を配置すると、より高い頻度および精度で位置検出が可能になる。モニタAPは、デフォルト設定では1.2秒間隔で各チャネルを監視する。

今回の実験では、図2.2に示すように全20台のAP(赤:通信AP 9台、青:モニタAP 11台)を配置した。モニタAPの設置により、APの総数はこれまで同規模で開催したWIDE合宿の倍以上が必要となった。部屋面積と面積あたりのAP数を表2.1、表2.2に示す。エリア全体の面積は図2.2における32m×77mの範囲(検出対象にしないトイレなどを含めた面積)としている。シスコ社が推奨する通信APあたりのカバレッジは約230-450㎡であり[51]、十分な通信APに加えてモニタAPを設置している。

2.3.2 無線LANのチャネル設計

2.4GHz帯においては、無線LANを利用する他の実験との住み分けのため11ch のみ利用した。今回のWIDE合宿では会場内全体を利用する無線LANの実験が本実験を含めて3つ平行して行われており、実験毎にエリア全体にAPを配置しているため、実験毎に1ch、6ch、11chが割り当てられている。実験毎にチャネルを割り当てた理由は、AP間の干渉が避けようがないため、同じ管理者の配下にあるAP同士が干渉したほうが、異なる管理者の配下にあるAP同士が干渉するよりも影響の把握や調整が容易であ

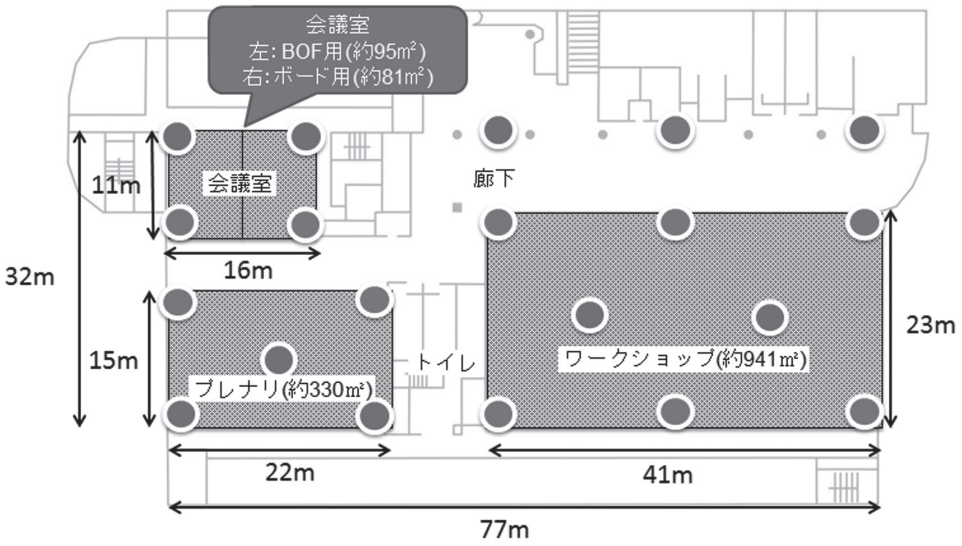


図2.2 実験エリア

るためである。

5GHz帯は、チャネルボンディングにより40Mhz帯域幅で運用し、チャネルはW53(52+56, 60+64)からW56(100+104,108+112,116+120,124+128,132+136)までを利用した。7チャネルが利用可能であるため、干渉が少ない運用が可能であった。まとめると、2.4GHz帯と5GHz帯の合計チャネル数は以下になる。

- 2.4GHz帯：計1チャネル(11ch)
- 5GHz帯：計7チャネル(52+56、60+64、100+104、108+112、116+120、124+128、132+136)

今回の実験では2.4GHz帯が1チャネルであるため、電波干渉により空間の通信キャパシティが小さく、また混雑すると端末から送信された信号を複数のAPで正しく受信できない可能性が上がる。そのため、可能な限り通信を5GHz帯に誘導すること、そして制御フレームによる2.4GHz帯の時間消費を抑制するチューニングを行った。

1）Band Select機能

Band Select機能は、APが2.4GHz帯において端末からのプローブを数回破棄することで、端末が5GHz帯で通信開始することを促す機能である。

2）接続レートの制限

下に示すように、2.4GHz帯では12Mbps以下の通信レートを無効に設定した。これにより、低速な通信や最低接続レートで送信されるAPからの制御フレームがチャネル時間を消費することを抑制した。5GHz帯では接続レートの制限は行わなかった。

802.11b/g Operational Rates

802.11b/g 1M Rate..... Disabled
802.11b/g 2M Rate..... Disabled
802.11b/g 5.5M Rate..... Disabled
802.11b/g 11M Rate..... Disabled
802.11g 6M Rate..... Disabled
802.11g 9M Rate..... Disabled
802.11g 12M Rate..... Disabled
802.11g 18M Rate..... Supported
802.11g 24M Rate..... Mandatory
802.11g 36M Rate..... Supported
802.11g 48M Rate..... Supported
802.11g 54M Rate..... Supported

802.11a Operational Rates

802.11a 6M Rate..... Mandatory
802.11a 9M Rate..... Supported

表2.1 部屋面積とAP数

場所	面積	通信 AP 数	モニタ AP 数	総 AP 数
プレナリ	330 ㎡	3	2	5 個
ワークショップ	941 ㎡	2	6	8 個
会議室(左)	95 ㎡	1	1	2 個
会議室(右)	81 ㎡	1	1	2 個
全体	2464 ㎡	9	1	20 個

表2.2 面積あたりのAP数

場所	面積/通信 AP 数（㎡/AP）	面積/モニタ AP 数（㎡/AP）	面積/総 AP 数（㎡/AP）
プレナリ	110	165	66
ワークショップ	470.5	156.8	117.625
会議室（左）	95	95	47.5
会議室（右）	81	81	40.5
全体	273.8	224	123.2

802.11a 12M Rate..... Mandatory
802.11a 18M Rate..... Supported
802.11a 24M Rate..... Mandatory
802.11a 36M Rate..... Supported
802.11a 48M Rate..... Supported
802.11a 54M Rate..... Supported

3）送信出力の手動調整

2.4GHz帯は11チャネル固定であり、APの送信出力は自動調整とした。5GHz帯では、はじめはAPの通信チャネルと送信出力を自動選択したが、APが密に設置されているため送信出力が低めに自動調整され、半分以上の端末がRSSが高い2.4GHz帯に接続する状態になったため、5GHz帯の送信出力を手動で設定し、APから5GHz帯の電波が端末に強く到達するように調整した。その結果、図2.3に示すように約7割の端末が5GHz帯に移行した。十数台設置されていたRaspberry Piを含めて5GHz帯に未対応の端末も存在するため、5GHzに対応した端末は大半が誘導できていたと思われる。

図2.4は合宿期間中の無線LAN端末数の推移である。ワークタイムは常時100台以上の端末が生活線として利用した。

図2.5と図2.6は5GHz帯のAP送信出力調整後のサイトサーベイ結果であり、RSSの分布を示している。サーベイ時には廊下の通信AP（1台）がトラブルによりオフラインの状態であったが、エリア全体で最低-60dBm以上のRSSが確保されており、2.4GHz帯よりも5GHz帯で強いRSSが確認できている。

2.4　データ分析

2.4.1　取得データの概要

上記のシステム環境により取得した位置データについて、検出の精度と傾向に関する分析を行った。分析には、MSEが提供するlocation history情報（以下、ヒストリ情報）を用いた。ヒストリ情報には、端末が初めて検出され

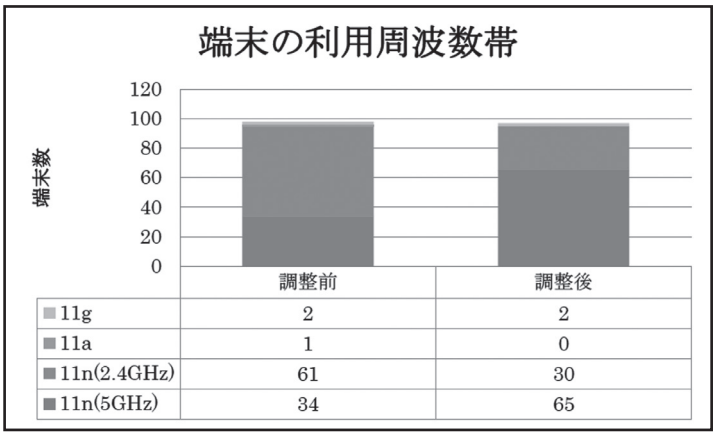


図2.3 端末の利用周波数帯（5GHz帯の電波出力調整前・調整後）

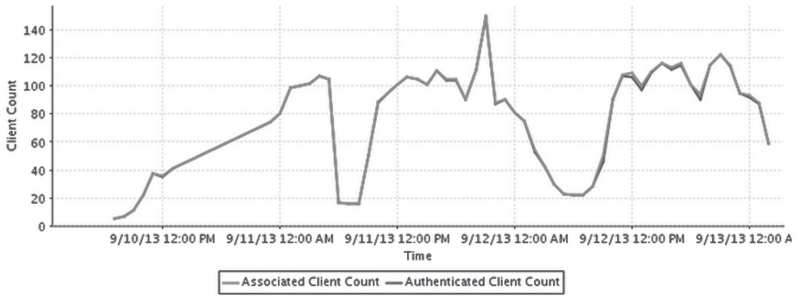


図2.4 接続端末数推移

たときと、端末が10m以上移動したと検出されたときに記録が残る。

本データの概要は以下の通りである。

- ・データ取得期間：2013年9月10日0:00～9月13日19:50
- ・検出した端末(mac address)数：1748
- ・サンプル数：95635

合宿参加者数129名に対して、macアドレスは1748検出されており、会場外から到来するシグナルも数多くデータに含まれていると考えられる。分析にあたり、まず検出精度が低い情報をconfidence factor値（以下CF値）により除外する。CFとは、ある端末の位置を計算した場合に、推定位置を中心とする正方形に端末が実際に位置することを95%信頼できるとした場合の、正方形の大きさを示す値である[52]。図2.7に示すように、CF値は信頼性95%の正方形における辺の半分の長さであり、x軸とy軸のどちらかで検出位置が実際の位置からの外れる距離のおよそ最大値と考えることができる。信頼性が高いデータは正方形の辺が短くなり（CF値が小さくなり）、信頼性が低いデータは辺が長くなる（CF値が大きくなる）。

図2.8に本実験のヒストリ情報におけるCF値の分布、図2.9にその値ごとの割合を示す。値が10m ～ 30mの範囲のサンプル数の割合が高く全体の46%を占めている。

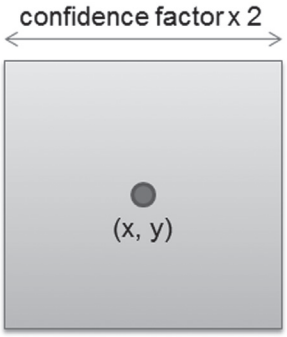


図2.7 confidence factor

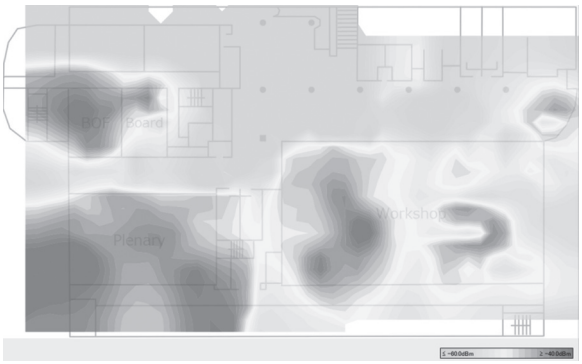


図2.5 2.4GHz帯のカバレッジ（RSS -40dBm ～ -60dBm）

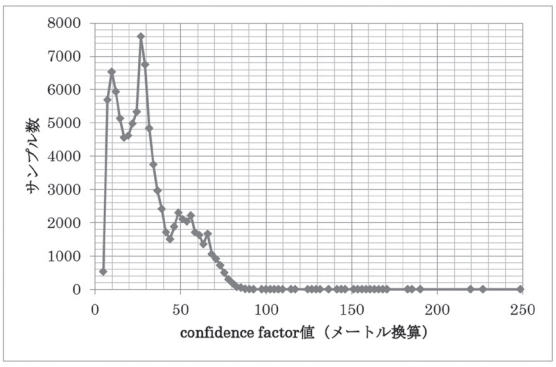


図2.8 CF値の分布

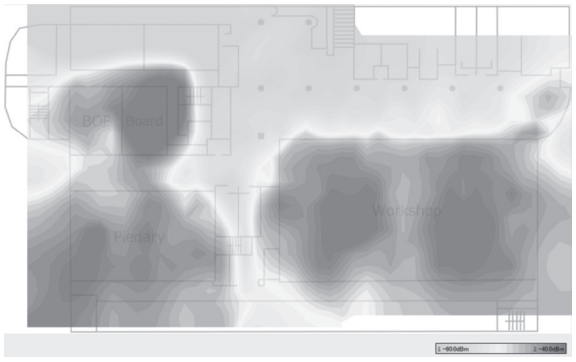


図2.6 5GHz帯のカバレッジ（RSS -40dBm ～ -60dBm）

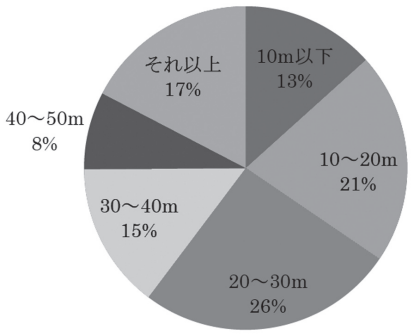


図2.9 CF値の割合

2.4.2　固定ノードの位置情報に関する考察

合宿地に固定設置され無線LANのネットワーク監視を行っていたRaspberry Pi端末について位置情報を分析した。図2.10 ～図2.19は、サンプル数が最も多かった5つのRaspberry Pi端末の推定位置である(単位:メートル)。各端末について、全サンプルの分布と、CF値30m以外にサンプルを絞り込んだ場合の分布を示している。検知位置が横方向に分散しているもの(固定端末1, 固定端末5)、密集しているもの(固定端末2, 固定端末3)、縦横に分散してしまっているもの(固定端末4)、など、端末毎に幾つかの傾向があるが、サンプルの密集度からノードの位置に依存して検出の精度にばらつきが生じていることが分かる。分散しているものほど推定位置が不正確と考えられる。このとき、同じ端末について、全サンプルとCF値30m以内のサンプルを比較しても大きな違いは見られない。そのため、CF値を利用して、精度が低いサンプルを完全に淘汰することは困難と思われる。

一方、検出位置が密集している(検出精度が高いと思われるエリアの)端末はCF値が30m以下のサンプルの割合が多く、位置が分散している(検出精度が低いと思われるエリアの)端末はその割合が低くなっている。例えば、比較的推定位置にばらつきが出ている固定端末1、4、5ではCF値が30m以内データは45.8%、74.8%、42.6%であるが、推定位置が密集している固定端末2、3では90%前後である。CF値のばらつきは、エリアの検出精度に関連すると考えられる。

2.4.3　位置による検出精度に関する考察

固定ノードの位置情報に関する考察より、ノードの位置により検出精度に差が生じていることが分かった。また、検出精度が悪い位置ではCF値が全体的に低下することが分かった。そこで、実験エリア内でどの場所が正確で、どの場所が不正確なのか検討する。

図2.20 ～ 2.22の3つの図は、CF値が10m以下、30m以下、30m以上のサンプルの分布である。見やすさのため、これらのデータは2013年9月12日の1日分としている。図2.20のように、精度が高い情報はモニタAPに周囲を囲まれたホール内や会議室内に集中しており、逆に精度が悪

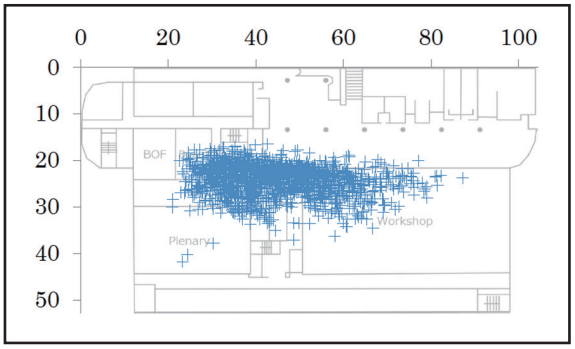


図2.10　固定端末1（全サンプル3142）

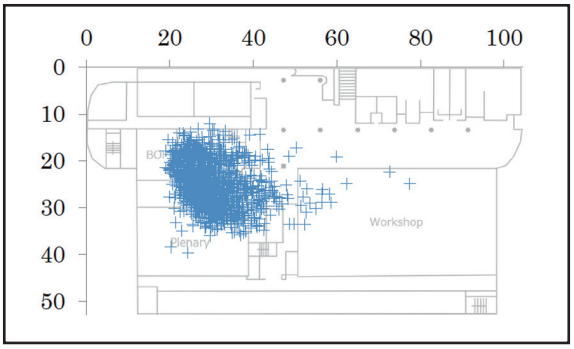


図2.12　固定端末2（全サンプル2703）

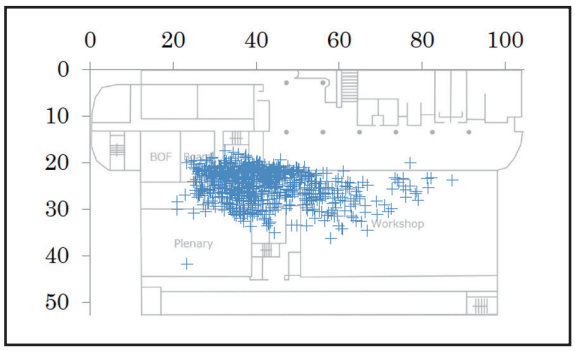


図2.11　固定端末1（CFが30m以内、45.8%）

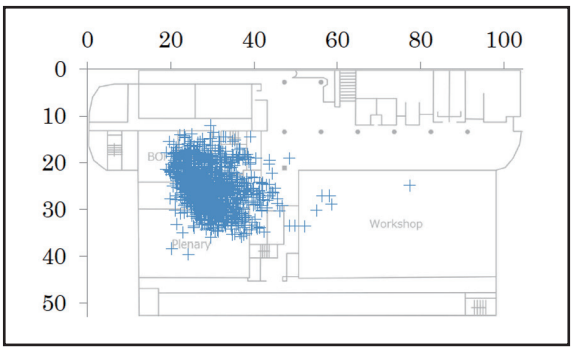


図2.13 固定端末2（CFが30m以内、89.7%）

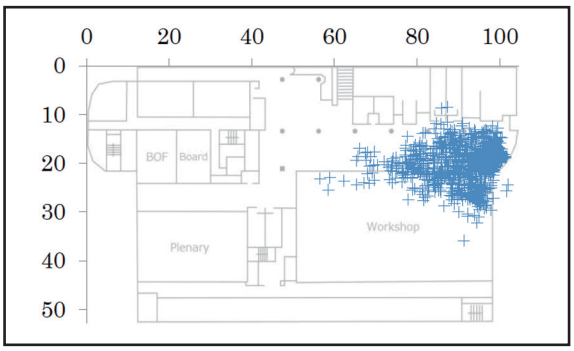


図2.14　固定端末3（サンプル数2694）

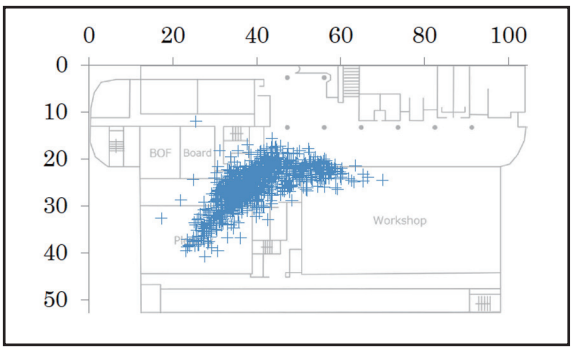


図2.17　固定端末4（CFが30m以内、74.8%）

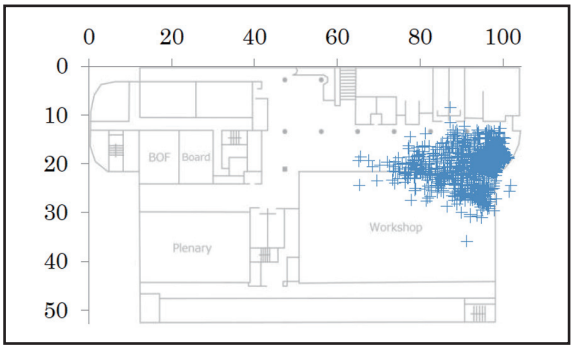


図2.15　固定端末3（CFが30m以内、91.3%）

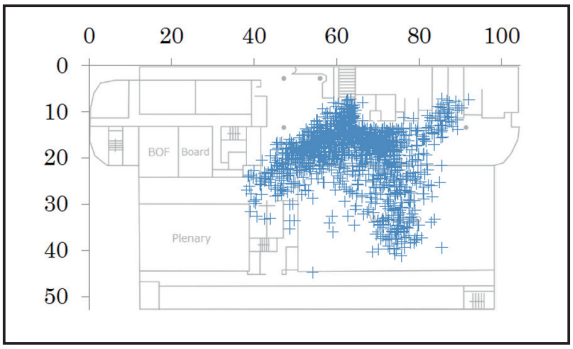


図2.18　固定端末5（サンプル数2669）

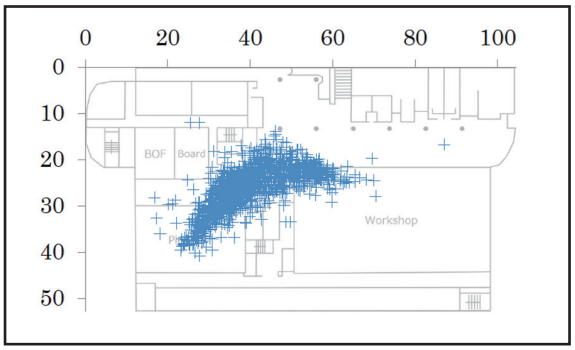


図2.16　固定端末4（サンプル数2676）

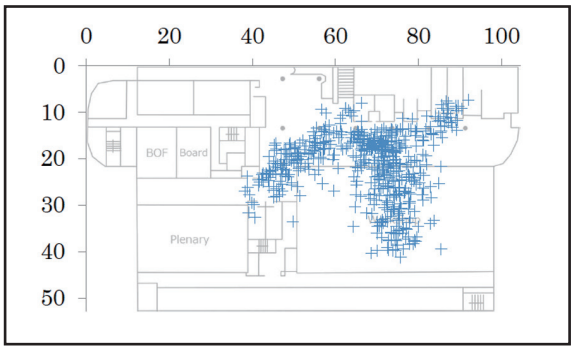


図2.19　固定端末5（CFが30m以内、42.6%）

表2.3　不特定端末と登録移動端末に関する位置情報

端末	全サンプル数	CF値が30m以内のサンプル数	CF値が30m以内のサンプル数の割合
不特定端末A	2149	355	16.5%
不特定端末B	2005	643	32.1%
不特定端末C	1343	349	26.0%
不特定端末D	1157	268	23.2%
不特定端末E	1125	520	46.2%
登録移動端末a	896	628	70.0%
登録移動端末b	586	389	66.4%
登録移動端末c	430	286	66.6%
登録移動端末d	392	292	74.5%
登録移動端末e	351	291	82.9%

い情報は図2.22のように中央の廊下のようにモニタAPに囲まれていない場所に集中している。廊下では、モニタAPと端末の間に壁や柱などの遮蔽物が存在することも検出精度に影響していると考えられる。

2.4.4　移動端末の履歴

次に、2.4.3で述べた固定端末以外で、検出された回数が多かった端末上位5台（不特定端末A ～ E）と、macアドレス調査により参加者の移動端末であることが判明している端末の上位5台（移動端末a ～ e）の分布を以下に示す。図示されているのは3日間のデータのうちCF値が30m以

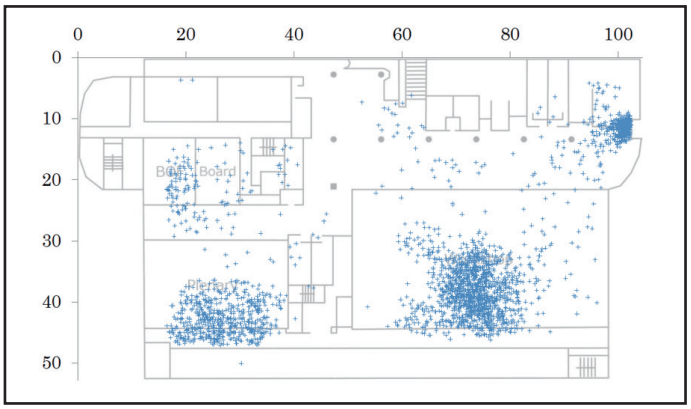


図2.20　CF値が10m以下のサンプル（サンプル数3234、全体の15.3%）

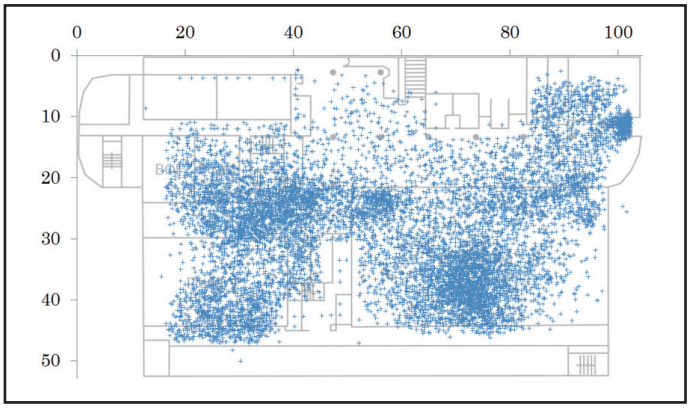


図2.21　CF値が30m以下のサンプル（サンプル数13769、全体の65.2%）

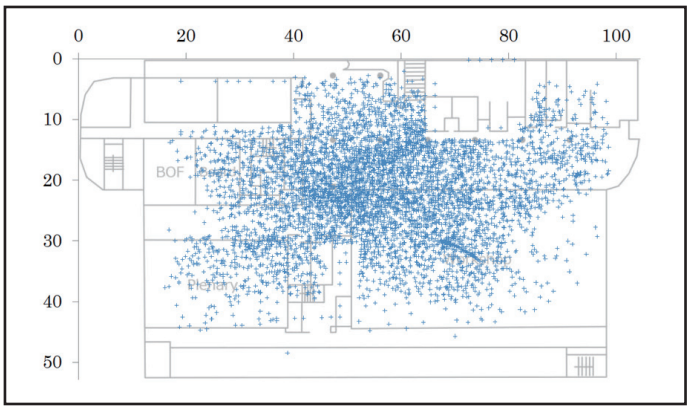


図2.22　CF値が30m以上のサンプル（サンプル数7338、全体の34.8%）

内のサンプルのみである。各端末の全サンプル数、CF値が30m以内のサンプル数、その割合を表2.3に示す。不特定端末は、A ～ Dのように図面の右上に履歴が集中する箇所があるが、表2.3に示すとおり精度が高いサンプル数の割合が低く、エリア外の端末を検出している可能性が高い。登録済端末は不特定端末よりも精度が高いサンプルの割合が多く、エリア内に現実的な値が取得できている。しかし、移動端末のデータは実際の利用と照らし合わせて考える必要があるため、今回は端末の実際の位置との比較はできなかった。移動端末に関するデータ精度の調査は今後の課題とする。

2.5 まとめ

WIDE合宿において、無線LANによるRTLSの実験運用を行った。初回の実験であるため、精度の検証は不十分であるが、以下のように構築・運用に関する有用な知見が得られた。

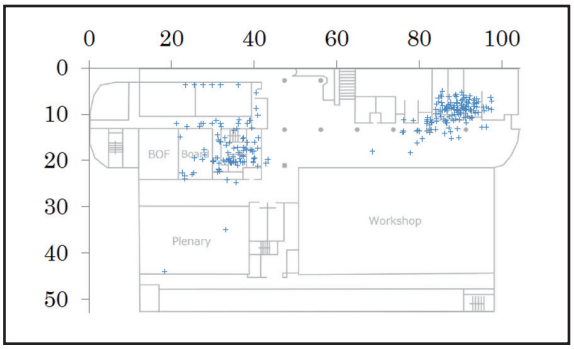


図2.23 不特定端末A

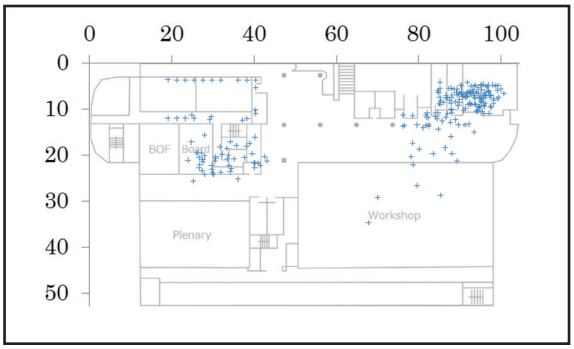


図2.26 不特定端末D

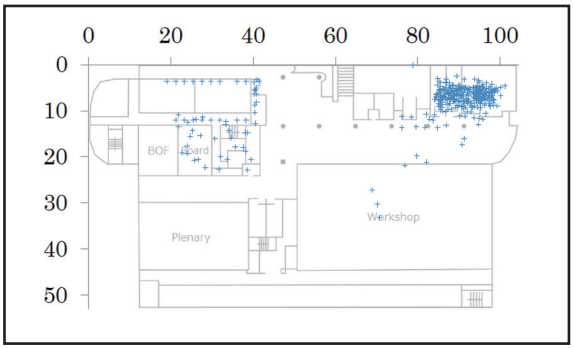


図2.24 不特定端末B

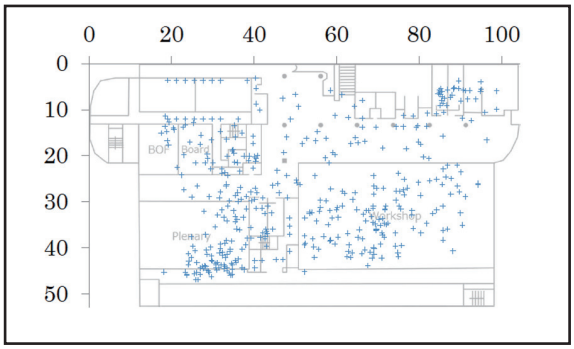


図2.27 不特定端末E

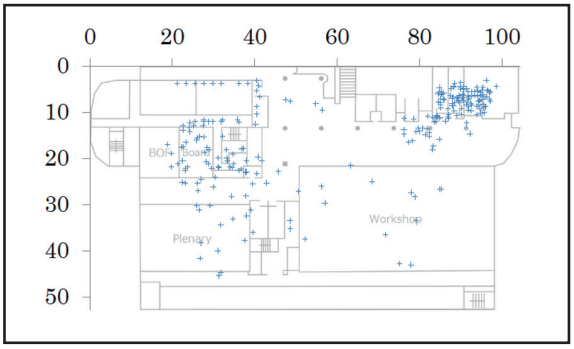


図2.25 不特定端末C

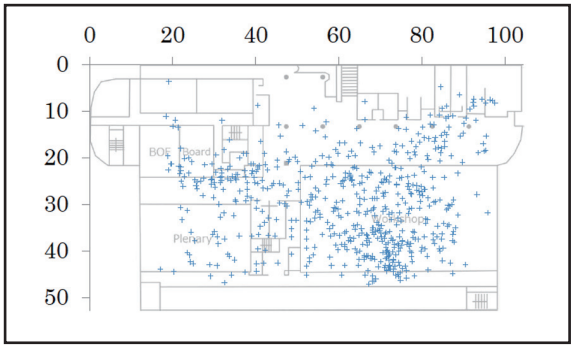


図2.28 登録移動端末a

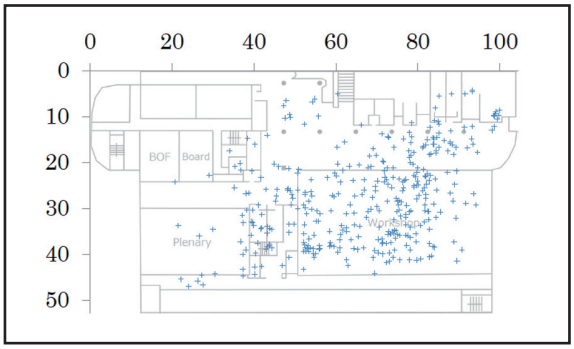


図2.29 登録移動端末b

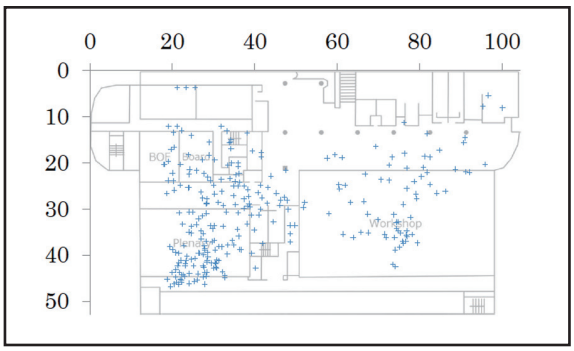


図2.30 登録移動端末c

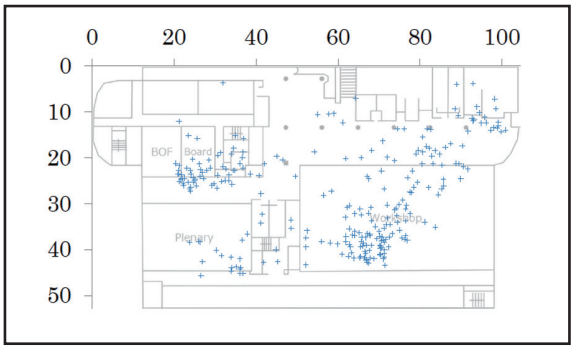


図2.31 登録移動端末d

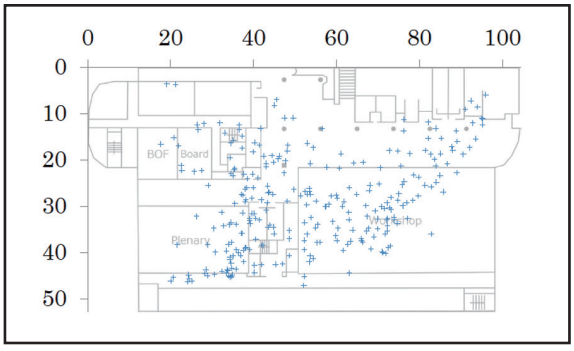


図2.32 登録移動端末e

・実験を行った合宿地は郊外であるが、予想以上に外来と思われる無線LAN端末が検知されている。外来の端末を効率的に除外する措置が必要である。

・本実験では、場所によって位置情報の精度に差が生じた。データはCF値によって信頼性が示されるが、その値のみで精度が悪いデータを淘汰することは限界がある。逆に検出精度が悪い場所では全般的にCF値が悪化するため、そのエリア内の位置精度が推測できる。

・位置精度の向上は、より綿密なモニタAPの配置と電波減衰に関するエリアの物理情報によって可能と思われる。しかし、それを複雑な実空間で簡易に取得・設計することは困難であり、位置情報の取得には既存の無線LANインフラよりも更にシステム管理者によるチューニングが必要である。

今後、位置情報精度を向上させるシステム構築・運用技術とデータ精度に関する検討を重ねると共に、このように簡易に取得できるようになった屋内位置情報のアプリケーション利用を進める予定である。

第３章　android端末用の室内型位置情報アプリケーションの開発と実験

3.1　はじめに

本章では、2013年9月9日から9月13日の5日間にわたって開催された「WIDEプロジェクト2013年9月合宿」において実施された「音を用いて位置情報を取得する実験」に関する報告を行う。

3.2　概要

今回行った実験は人間の可聴域外の音を出すスピーカーとAndroid端末を用いて人の位置情報を取得する実験である。GPSや電波では正確に位置の特定をすることが困難な屋内での位置情報取得方法を提案した。この実験は、インテック社が開発したライブラリを利用し、サーバの構築とクライアントのソフトウェア(Ebaran-doroid)を作成することによって実現した。

3.3　システムの構成

3.3.1　システムの構成

今回のシステムは全体的にはサーバ、クライアントで構成されている。クライアントのandroid端末には今回開発したEbaran-droidをインストールし、必要事項(UUID, macアドレス, 名前, 時間, 位置情報, アンケート項目１～６(IPv6, 情報セキュリティ , クラウド, 可視化, 新人, 学生),コメント)を入力、サーバ側に送信する。その後、アプリを表に出していて、かつスピーカーから認識されている場合、随時位置情報が送信される。サーバはandroid端末から送られた情報をデータベースに保持し、それらを地図上に書き出す機能を持っている。集めたデータはphpmyadminにログインすることで閲覧が出来る。

また、自身の端末の登録は出来ないが、閲覧のみであれば一般的なウェブブラウザでのマップ画面の閲覧が可能である。

3.3.2　実装

今回の実験で使ったソフトとハードのシステムを図3.1に示す。クライアントは一般的なandroid端末であり、サーバの特定のURLにアクセスすると図3.2のような画面が表示される。色線で囲まれた部分をクリックすると、現在その場所に居る人の詳細な情報を閲覧することが出来る。

3.4　実験

3.4.1　場所

今回の実験場所は信州松代ロイヤルホテルである。会場の図を図3.3に示す。今回スピーカーを13台用意し、喫煙所に１台、ワークショップルームに６台、プレナリールームに４台、BoF部屋とボード部屋に各１台ずつ設置をした。図3.3の矢印はスピーカーの設置されている位置と音の向いている方角である。

3.4.2　結果

今回の合宿では、音声を用いて位置情報等のデータの取得が出来ることを確認できた。今回、データベースはMySQLを使用しており、9月13日正午確認でMacアドレス数50件、位置情報履歴(図3.4 locationテーブル) 371件を取得することが出来た。集めたデータはwide証明書による認証をすることで、データ解析としての利用ができ、PC企画の”データ解析ごっこ”で利用された。

任意のアンケートでは、ユーザインターフェイスの面ではアプリケーションの利用者から様々な指摘を受けた。(図3.5)なお、このアンケートの一部は合宿期間中に修正済みの点もある。その中でも特に要望が多かったものが”今、自分がスピーカーから認識されているか”であった。

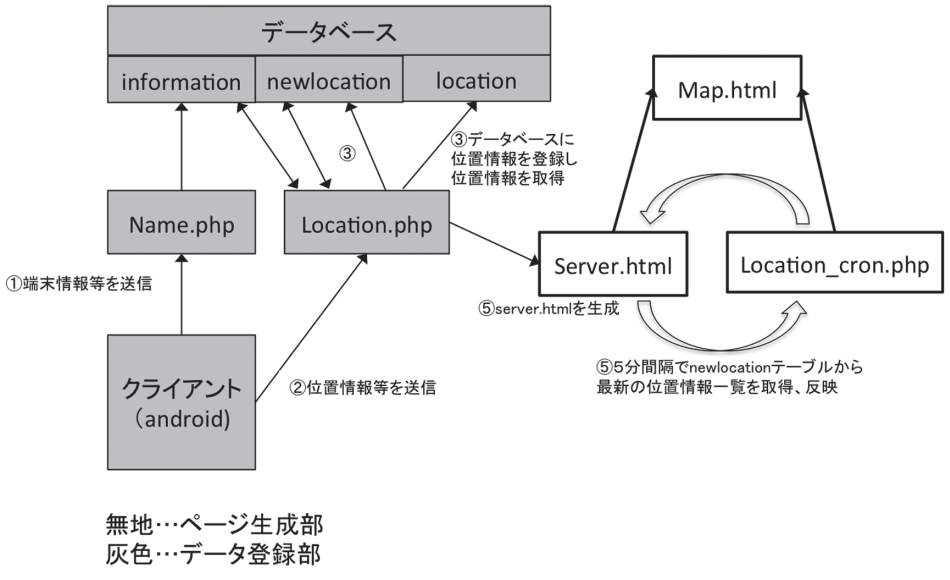


図3.1　システム仕様

3.5 問題点

指向性がある為(スピーカーを中心に上下左右３０度)認識されるようにスピーカーの前にわざわざ行かなければならない場面もあった。人間には聞こえない高周波の音でも、聞こえる人や、実際には聞こえていないにもかかわらず、スピーカーがあることで、聞こえると思い込む人がいた。

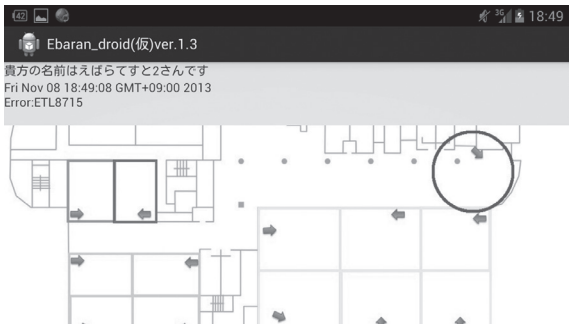


図3.2 android端末でアクセス時の画面(map.html)

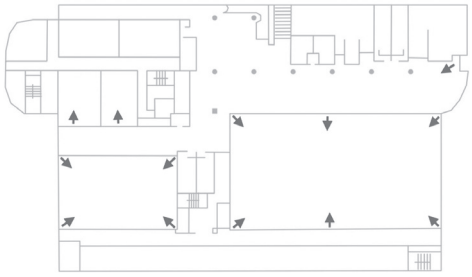


図3.3 会場図とスピーカー設置位置

id	macaddress	datetime	locationAP
18b557d2-afc5-49c7-9fe2-b4df6d44297d	10:bf:48:ca:08:b9	2013-09-10 10:42:52	1
18b557d2-afc5-49c7-9fe2-b4df6d44297d	10:bf:48:ca:08:b9	2013-09-10 10:42:52	1
18b557d2-afc5-49c7-9fe2-b4df6d44297d	10:bf:48:ca:08:b9	2013-09-10 10:45:18	130710091
18b557d2-afc5-49c7-9fe2-b4df6d44297d	10:bf:48:ca:08:b9	2013-09-10 10:45:18	130710091
111111	123456	2013-09-10 10:48:50	130710065

図3.4 取得データの表示

感想/Comment

マイクの場所も図で示して近づくように知らせたほうがいいのではないかと思います。 字が読めない。 マップ画面で、部屋に人数が表示されたりするといいかも。 また、クリックしないと一覧がみえないのは不便な気がするので、画面に表示しきれる分は表示してしまうといいかも。 クリックしてポップアップもいいけど、マップの下の方に詳細情報divとかを作って、クリックするとそこに一覧を表示するとかするといいかも。 リンゴユーザーに愛の手を ポップアップが押すたびに下がっていって見えなくなるのが困ります 精度が上がったら、探してる人の方向を常に示してくれるコンパスとかつけてください。待ち合わせアプリ的な アプリに、探してる人を入力すると、プッシュ通知が行って、許可を得れば位置情報を返してくれる機能をつけたい バックグラウンド動作 自動更新 エリア内人数表示 自分の履歴確認 が欲しいです アプリから登録されないみたい・・・ 今音が聞こえていかどうか。つまり自分の位置がわかっているのか、わかるようにして欲しいです 自分のいる位置を動的に表示してくれるともっと嬉しい

図3.5 アンケート自由コメント欄

第４章　ウェブトラッキング可視化実験報告

4.1 はじめに

ウェブトラッキング[55, 57, 59, 56]は，ターゲティング広告やトレンド解析などを行うための手法であるが，これはユーザのウェブ閲覧履歴を密やかに収集しているため，深刻なプライバシ問題となっている．そこで，誰がどういった情報を集めているかをユーザに明確に見せる，ウェブトラッキング可視化システムの開発を行った．本システムの有効性を確認すべく，2013年秋のWIDE合宿にて実験を行ったため，ここに結果を報告する．

4.2 提案システム概要

本システムはユーザのトラフィックをキャプチャし，アプリケーションレベルの解析を行うことで，ウェブトラッキングの可視化を行う．アプリケーションレベルプロトコルの解析には，現在，開発を行っているCate-naccio DPI(CDPI)[54]を利用した．CDPIはpcap[60]を利用してトラフィックをキャプチャし，アプリケーションプロトコルのパースを行った後，結果をMongoDB[58]に保存する．データの解析にはMongoDBの提供している，MapReduce機能を利用して行った．また，最終的な可視化はPythonを用いてHTMLファイルと，Graphvizの.dotファイル，Cytoscpaeの.sifファイルを出力し用意したウェブサーバを通して，参加者に提供した．

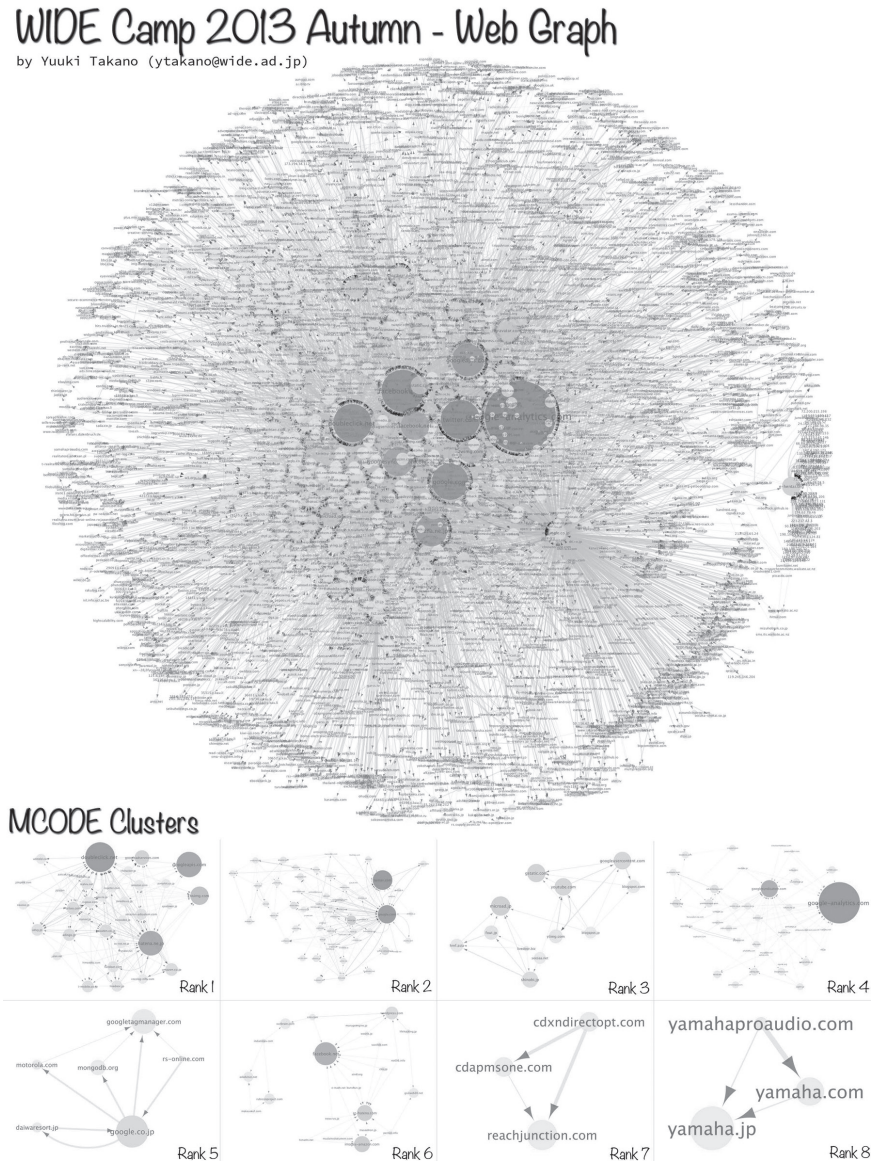
4.3 実験結果とまとめ

表4.1は2013 年秋のWIDE 合宿参加者の内訳である．本実験では，これらの参加者が合宿中に利用したネットワークトラフィック全てをキャプチャし解析し，得られたHTTPリクエストの数は合計で734,194リクエストとなる．図

4.1は，取得したHTTPリクエストのリファラヘッダから，ウェブの参照状態を可視化したものである．ここでは，例えば，platform0.twitter.comとplatform.twitter.comの2つがtwitter.comとなるようなドメイン集約を行い可視化を行っており，その結果，3,966個のノードと12,941個のエッジとなった．

本システムでは，被参照数が多いほど，ウェブトラッキングを行っている可能性が高いと推定して可視化している．そのため，図4.1では，入ってくるリンクの数が多いノードほど大きく表示している．今回取得できたデータでは，表4.2で示されるサイトが，最も多く参照されていることがわかった．これらは，広告，及びソーシャルサイトであり，実際にウェブトラッキングを行っている事が知られている[59]．

また，図4.1の下部で，MCODE[53]を用いたクラスタリングを行った結果を示している．その結果，Rank1のク



https://github.com/ytakano/pictures/blob/master/webtracking/wide1309.png

図4.1 WIDE 合宿参加者のウェブ閲覧グラフ

ラスタに下記で示す広告サイトが多く含まれている事が判明した.

doubleclick.net, amazon-adsystem.com, googleadservices.com, i-mobile.co.jp, advg.jp, adingo.jp, iogous.com, admeld.com, criteo.com

同様に, 他のクラスタにも多くの広告サイトが含まれていることが明らかとなった. 被参照数による方法では, Facebook等の有名なサイトしか抽出することが出来なかったが, クラスタリングを行うことで, 多数の広告サイトを抽出することが可能となることが本実験で得られたデータを解析することによって明らかとなった.

本実験で, 提案システムが実際にウェブトラッキングを行っているサイトを可視化可能であることが, 実ユーザのデータを用いて証明することが出来た. また, 詳細なデータ解析を行うことで, クラスタリングが広告サイトの抽出に有効であることが判明した. 今後は, 実験から得られた知見を元にシステムの改良を行い, 引き続きWIDE合宿にて実験を行っていきたい.

表4.1 2013 年秋のWIDE 合宿参加者

	社会人	学生	合計
男性	78	39	117
女性	1	11	12
合計	79	50	129

表4.2 最も参照されている上位5 サイト

サイト	被参照数
google-analytics.com	847
facebook.com	437
twitter.com	393
doubleclick.net	380
google.com	356

第５章 ネットワークトラフィックのL7解析によるユーザ別ウェブページ閲覧履歴の取得

5.1 実験背景

情報技術の発展によって, ネットワークを使った様々なサービスが展開されている. そして近年のビッグデータという言葉に代表されるように, 大量の情報を解析し, 規則性やユーザの特徴付けをおこなうことにより, さらに付加価値の高いサービスを提供しようとする気運が高まっている. このような世情の中で, 高速なスループットの獲得に専念してきた従来のネットワークインフラストラクチャに対し, 慶應義塾大学西宏章研究室(以下西研究室)では, 新たな一面としてユーザがネットワークを流れる情報を自由に扱うことのできる次世代インフラストラクチャの開発を目的とした研究活動をおこなっている. その構成要素の根幹となるサービス指向ルータは, 従来のパケット転送と同時にネットワークL7情報を抽出し, データベースに蓄積する. このルータにより, ネットワークトラフィックの全レイヤ情報の融合が可能となり, 新たなサービスの実現やより効率の良いネットワークの利用が期待される. 例えば, 現在利用提供されるサービスの行動履歴をプロバイダに問わず統合し, ユーザの嗜好を従来よりも細かく反映したリコメンデーションサービスなどが考えられる.

本実験では, 次世代インフラストラクチャ実現に向けて開発されたネットワークトラフィックからL7情報を抽出するソフトウェアNEGIを利用し, WIDE CAMP 2013 AutumnでのL7解析をおこなった. 複数のプロバイダの提供するサービスの横断的な取得を想定し, ウェブページ閲覧をターゲットとしたユーザ別の閲覧履歴を取得した. そして取得した情報を利用し, ウェブページのリコメンデーションをおこなった. 本報告書では, その取得方法と実験結果, 閲覧履歴の活用方法について言及する.

5.2 L7アナライザ: NEGI

サービス指向ルータの実現にあたって, ネットワークを流れる情報をルータ上にて取得し, サービスに利用する部分を抽出する必要がある. 西研究室では, 上記の要件を満たすオープンソース・ソフトウェアNEGI (Negi

Enables Great Intelligence) を開発している．具体的な処理としては，ルータ上にてパケットからTCPストリームの再構築をし，続いてアプリケーションレイヤで施されるHTTP/1.1(gzip, chunk)エンコードのデコードをおこなう．デコードしたペイロードに対し，指定した文字列で検索をかけることにより，所望の文字列を得る．

NEGIの実行結果の例を表5.1に示す．パケットヘッダ情報に加え，指定した検索文字列の後に続くペイロードを保存している．閲覧履歴の取得はこの情報を用いておこなう．

5.3 ユーザ別ウェブページ閲覧履歴の取得

NEGIによってL7解析された情報を用いて，ユーザ別ウェブページ閲覧履歴を取得する．本実験において，閲覧履歴取得のためにNEGIに設定した文字列はHTTPヘッダの“GET”，“Host:”，HTMLタグの“<title”である．まず，同一のストリームに属する“GET”，“Host:”の後に続く文字列を連結することにより，URLを得る．ウェブページのタイトルは”<title>”タグで囲まれた文字列から抽出する．“GET”や“Host:”のHTTPリクエストヘッダとHTMLタグは通信方向が異なるため，URLとタイトルはIPアドレスとポート番号の入れ替わりによって関連付けられる．

5.4 閲覧履歴取得実験

実験環境を図5.1に示す．WIDE CAMP 2013 Autumnにおいて，参加者が利用する生活線をミラーリングし，Analyzing Serverに流した．Analyzing Serverのイーサポートに対し，NEGIを実行することでトラフィック解析をおこなった．抽出情報はPostgreSQLに保存され，このデータベースに対しウェブページ閲覧履歴取得のプログラムを実行することにより，合宿参加者の閲覧履歴を得た．

5.5 実験結果

CAMP中に実施したアプリケーションの実行結果を図5.2，5.3に示す．図5.2はウェブページのタイトルを抽出し，タイムスタンプ，MACアドレス，送信元，宛先IPアドレスと共にタイムスタンプが新しい順に表示している．ユーザが新たにウェブページを閲覧したならば，NEGIによるL7解析によりデータベースが更新され，情報が即座に結果に反映される．図5.3はユーザ別の閲覧履歴を示している．ユーザ識別のキーとなるMACアドレスを入力す

表5.1 NEGI実行結果の例

ID	Pattern	Result
1	GET	/ HTTP/1.1
2	Host:	www.yahoo.co.jp
3	<title	>Yahoo! JAPAN<

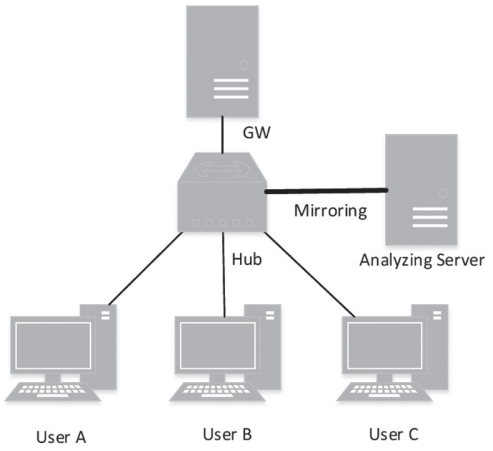


図5.1 実験環境

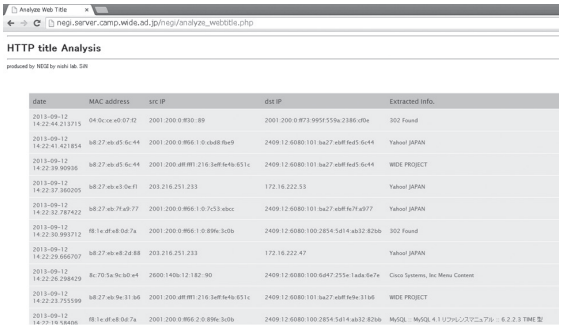


図5.2 HTTPタイトル解析アプリケーション

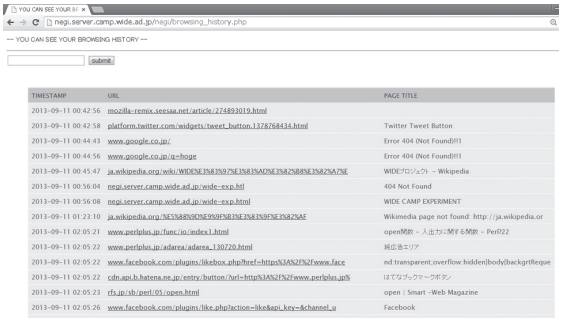


図5.3 ユーザ別閲覧履歴取得アプリケーション

るとCAMP中に閲覧したすべてのウェブページのURLをタイムスタンプ，タイトルと共に表示する．いずれの場合においても流れるパケットを解析するため，ユーザ側で新たなアプリケーション等を準備せずとも，情報収集が可能である．

5.6 滞在時間情報を利用したリコメンデーション

リコメンデーションはAmazon社がおこなう「この商品を買った人はこんな商品も買っています」として，ユーザの嗜好に合わせた関連商品を提示するサービスに代表される手法である．効果的に商品の推薦をおこなうことにより，プロバイダ側としては購買率の向上が期待され，ユーザ側としては短期間により多くの好みの商品を閲覧することができるという利点がある．

取得したユーザ別閲覧履歴を用いて，ユーザへのウェブページリコメンデーションをおこなう．プロファイリングのための指標として，ウェブページ滞在時間を用いた．この滞在時間は連続したウェブページのリクエストタイムスタンプの差分によって算出している．タイムアウトを30分に設定し，タイムアウト時間を経過したページは既に離脱済みのページと判断して，リコメンデーションのための情報からは除外する．長く滞在したウェブページほど，興味が高いと考え，算出した滞在時間に応じて，1から3のユーザのウェブページに対する評価値を割り振る．評価値決定のためのモデルとしてワイブル分布を用いた．このモデルは一定時間経過時のウェブページ離脱確率を表している．この確率分布関数を3つのエリアに分割し，エリアに応じて1から3の評価値を決定した（図5.4）．CAMP初日のユーザ閲覧履歴から算出した滞在時間を教師データとして，ユーザごとにモデルを構築し滞在時間を評価値に変換した．この処理によって生成

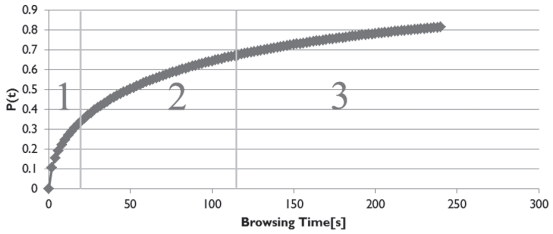


図5.4　ワイブル分布を用いた評価値算出例

したユーザ - ウェブページの評価値行列を用いて，協調フィルタリングをおこなう．まずピアソン相関によってリコメンデーションを受けるユーザ（活動ユーザ）と他のユーザとの類似度を算出し，その類似度を利用した加重平均によって活動ユーザの閲覧していないウェブページの予測評価値を求める．算出した予測評価値の高いウェブページほど活動ユーザが好むことが予想されるため，そのウェブページを出力として活動ユーザに推薦する．具体的には以下の(1), (2)式にて活動ユーザaの未閲覧のウェブページjに対する予測評価値$\hat{s}_{aj}$を算出した．

$$\rho_{ai} = \frac{\sum_{k \in Y_{ai}} (s_{ak} - \bar{s}'_a)(s_{ik} - \bar{s}'_i)}{\sqrt{\sum_{k \in Y_{ai}} (s_{ak} - \bar{s}'_a)^2} \sqrt{\sum_{k \in Y_{ai}} (s_{ik} - \bar{s}'_i)^2}} \quad (1)$$

$$\hat{s}_{aj} = \bar{s}_a + \frac{\sum_{i \in X_j} \rho_{ai} (s_{ij} - \bar{s}'_i)}{\sum_{i \in X_j} |\rho_{ai}|} \quad (2)$$

このとき，ρ_{ai}はユーザaとユーザiのピアソン相関，Y_{ai}はユーザaとユーザiが共通に閲覧しているウェブページの集合，X_jはユーザの集合である．

従来の協調フィルタリングはユーザがアイテム（本実験ではウェブページ）を評価した数が少ないと正確な推薦ができないという欠点があるが，滞在時間という暗示的な指標を適応することにより，この問題に対応できると考えている．また，既存のオンラインショッピングサイトや動画配信サイトにて，リコメンデーションのために利用される情報は，そのサイトでの利用者の行動履歴という制限があるが，ネットワークトラフィックのL7解析ならばその制限が解除され，サイト横断的に収集した情報に基づくリコメンデーションが実現可能である．

5.7 結論

次世代インフラストラクチャ構築に向けてのネットワークトラフィック解析ソフトウェアNEGIを100人規模が利用するネットワークにおいて適用し，ユーザ別ウェブページ閲覧履歴を取得した．取得した情報を利用して，ウェブページのリコメンデーションをおこなった．

第6章　まとめ

2013年9月の合宿においては、データ解析に関する検討も行ったが短時間での実験であったため十分な検証はこれからである。今後しばらく合宿において同様の情報収集及び解析を進める予定である。また、社会受容性の検討に関しても並行して進め合宿だけでなく広く情報を収集し活用するプラットフォームへの展開を進める。

なお、収集されたデータはWIDE Project内において活用できるように準備を進めている。これはBigdataに関するWGを設立しその中で進めて行く予定である。