簡易センサを用いた短周期観測データの継続的アーカイブとオープンデータ化に関する研究

福井大学地域環境研究教育センター研究紀要
「日本海地域の自然と環境」
No. 27, 97 - 104, 2020

簡易センサを用いた短周期観測データの継続的アーカイブとオープンデータ化に関する研究

A case study on long-term archiving and real time online retrieval of short-periodic observation data from simple IoT sensors

浅越　陽一*
（福井大学　工学部　情報･メディア工学科）
周　　睿**
（短期留学プログラム学生［蘭州交通大学］）
福間　慎治***
（福井大学　学術研究院工学系研究部門
情報･メディア工学分野）
森　眞一郎****
（福井大学　学術研究院工学系研究部門
情報･メディア工学分野）

１．はじめに

現在，様々な環境計測データがオープンデータとして入手可能となりつつあり，気象系のデータとしては 10 分毎の水位データや平均風速等のデータが入手可能となってきた．我々の研究室でも，本センターの過去の助成等を契機として試作した簡易風速計測システムを工学部３号館４階に設置し，（停電や機器故障などの期間を除き）約３年間の１秒単位の風速データを蓄積してきた [1]．気象や気候のような中長期的，かつ，比較的広域を対象とした現象の把握には全く無意味とも思えるデータであるが，この種の観測データを多数集約できれば，局所的かつ極めて短時間に発生する気象現象の把握などへの活用の可能性が広がる．本論文では，このようなデータを長期にわたり継続してアーカイブし，かつ，一般ユーザに対して公開することを目的として実施した，長期保存データに対する検索システムの高速化とデータ公開のための Web インタフェースの開発について報告する．

２．研究背景

２．１　簡易センサを用いたリアルタイム環境計測システム

我々の研究室では平成 28 年度の地域環境研究センター研究費支援「オープンデータと連携した簡易シミュレーションによる地域環境の実時間予測に関する研究」の一環として一秒毎の風速と照度を計測する簡易環境計測システムの開発を行った．安価で簡易なセンサを多数もちいたセンサーネットワークの構築に関する課題の探求と１秒毎の短周期観測データを遠隔地のクラウドサーバに集約する際の実環境における課題の探求がシステム開発の目的であった．次世代通信規格５Ｇにおけるリアルタイム性と大容量通信の特性を利用するアプリケーションの一例とも考えることができる．

安価で簡易な計測システムの構築という安価なソーラーパネル（100 円均一ショップで購入した園芸用照明を分解して使用）と風杯型風速計 [2] の保守用風杯 (約 2000 円) をセンサデバイスとして使

（キーワード：環境センシング，オンラインアーカイブ，風速，福井）

* Youichi Asakoshi（Division of Information Science, Faculty of Engineering, University of Fukui, Fukui, 910-8507）
** Zhou Rui（UFSEP Visiting Student from Lanzhou Jiaotong University, China）
*** Shinji Fukuma, **** Shin-ichiro Mori
（Department of Information Science, Graduate School of Engineering, University of Fukui, Fukui, 910-8507）

用し，前者はソーラーパネルの発電電圧から照度を，後者は風杯の回転に伴い ON/OFF を繰り返すリードスイッチの ON/OFF 周期から風速を算出する．電圧や周期を照度や風速に換算し，遠隔地のサーバに送信する機能デバイスとしては Raspberry Pi2 B を用いてシステムを構築した．風杯は本学工学部 3 号館 4 階の非常階段に設置し Raspberry Pi2 B との間は RJ11 ケーブル（通称、電話線）を用いて接続した．ソーラーパネルは室内の Raspberry Pi2 B に A/D コンバータを介して直結した．センサーノードで計測されたデータはクラウドサーバ（Amazon EC2 東京地域）上のデータベースに蓄積するとともに，一部のデータは Web インタフェースを介して検索可能としている．この際，センサノードとクラウドサーバとの間の通信は SSH Port Forwarding を用いることで，悪意のユーザからの不正データ挿入を回避している (図 -1 参照)．なお，令和元年度の本センター助成を受けて，当該システムの改良（1: 風杯と機能デバイス間を ZigBee 規格の無線接続に変更することで可搬性を改善，2: 4 つのセンサノードからの情報を集約・集計して遠隔地のサーバに送信する機能の追加，3: 機能デバイスの小型化、省電力化）をおこなっている [3].

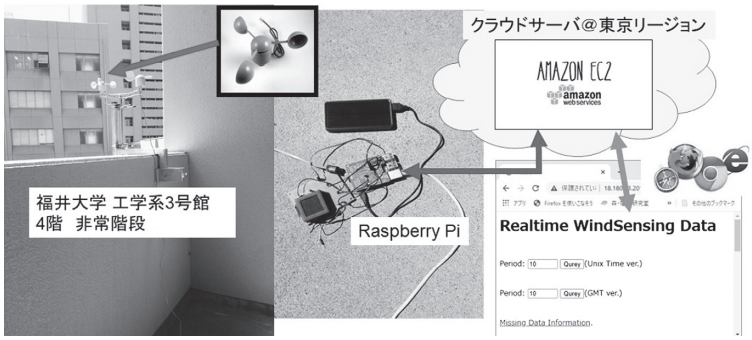

図-1　既存システムの概要

観測データを格納するデータベースとしては，Raspberry Pi 等の軽量デバイスでも動作が可能な Sqlite3 を用いて実装している．大手クラウドサービスベンダが IoT デバイス向けのデータ集約サービスを提供しているが，軽量デバイスには負荷の高いユーザ認証や高機能であるが複雑な API の利用が必要であり，通信データ量に対する課金の問題もあることからこれらの利用は見送った．

クラウドサーバに蓄積されたデータに対しては Web ベースの簡易なユーザインタフェースを介して検索が可能である．しかしながら，データ量の増加とともに，検索時間が増加するとともに検索のためのデータベースアクセスとデータ登録のためのアクセス競合によるデータ欠損の問題が発生した．そのため，検索対象範囲を直近数カ月に限ることでこれらの問題を回避した．しかし、その代償として検索対象範囲外の過去のデータへのアクセスができなくなってしまった．そこで本研究では，これらの問題を解決すべくデータベース検索の高速化と継続的なアーカイブを可能にするとともに，検索の利便性を向上する Web インタフェースの開発を行った．

3．Web ベース検索システムの高速化

3．1　検索アルゴリズムの改良

先行研究 [1] のシステムでは，検索インタフェースとして検索依頼がデータベースに到着した時刻の直前 N 個分のデータを検索・提示する機能を用意していた．この際，データベースへのデータ格納順が時刻順であることを完全に保証できないシステムで多用されている時刻逆順ソート後に先頭 N 個のみ出力するプログラムを実装していた．そのため，検索毎にデータベース全体に跨るソートが発生しオーバーヘッドの主要因となっていた．特に数ヶ月分のデータが格納されたデータベースでの検索では数分単位の待ち時間とり，実時間検索とは言い難い状況が発生していた．

しかしながら，我々のシステムでは，センサノード数が増加すると局所的な時刻順の乱れは発生するもののその影響は限定的であり，データベース全体に跨るソートは必須ではない．そこで，大域的な逆順ソートを省略し，注目領域に限定した順ソートを行うことで大幅な速度向上を目指すこととした．この際，注目領域の指定方式として，検索範囲の開始時刻 A と終了時刻 B を指定してその範囲内のデータを抽出する方式（以下 Between 方式と呼ぶ），ならびに，検索範囲の開始時刻

のデータベース先頭からの Offset と表示するデータ数 C (理論上は C=B-A) を指定して検索する方式 (以下 Offset 方式と呼ぶ) の検討を行った．また，従来方式の自由度を上げ逆順ソート後に Offset 方式を適用して任意時刻からの検索を可能にする方式（以下 Sort(D) 方式と呼ぶ）を実装し，これら 3 種の検索アルゴリズム（図 2 参照）の比較実験を行った．

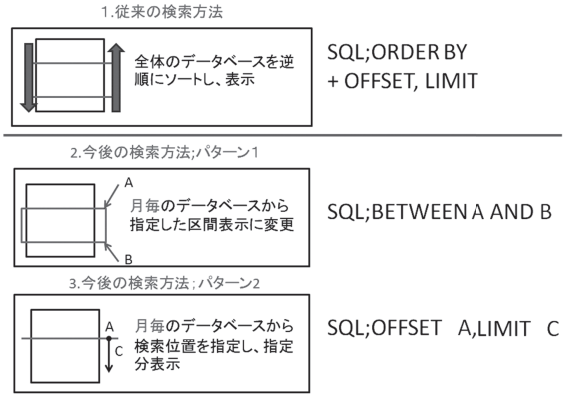

図-2　3種の検索アルゴリズム

2018 年 4 月の 1 ヶ月分のデータ（38.5MB）と 2018 年 3 月 28 日から 7 月 27 日までの約 4 か月分のデータ（151MB）から，2018 年 4 月の 1 日，15 日，ならびに 30 日の 0 時 0 分 0 秒から 100 個分のデータを 3 種類の検索方式を用いて検索した場合の実行時間を表 2 に示す．

表-2 検索アルゴリズム毎の検索時間[s]

検索対象日	1 ヶ月分データ			4 ヶ月分データ		
	Sort(D) 方式	Between 方式	Offset 方式	Sort(D) 方式	Between 方式	Offset 方式
4/1	28.9	0.68	0.02	121.3	2.74	0.02
4/15	30.7	0.86	0.24	124.4	2.91	0.29
4/20	12.5	1.05	0.48	123.2	3.09	0.52

Sort(D) に比べて他の 2 方式が圧倒的に高速化できることが確認できた．また，ソーティングは行わないがデータベース内を開始時刻と終了時刻の両方で検索する Between 方式に比べて，開始時刻に相当するデータベース内の開始位置を明示的に指定する Offset 方式は更に高速であるとともに，データベースサイズの影響を受けにくいことが確認できた．

しかしながら，Offset 方式では，検索結果が検索範囲と一致しない現象が確認できた．これは，あらかじめ予測ができない観測データ欠損によりデータベース内の実際の Offset 位置が理論上の Offset と異なってしまうことに起因したものであり，長期間のデータを格納したデータベースで大きなずれとなるだけでなく，データベースのデータ領域外への不正アクセスを行ってしまう問題も発生する．そこで，次節ではデータ欠損にともなう各種問題を回避する手法について検討を行う．

3．2　データ欠損に対応した局所的ソート

IoT デバイスからの観測情報の収集では，ネットワークの輻輳やデータベースのアクセス集中など事前に予測不可能な短期間のデータ欠損についてシステムレベルでの検討が必要である．高信頼性が求められる観測システムでは，相応のコストを払うことでデータ欠損の回避や軽減が図られるが，本研究で対象とするような簡易センサを用いた観測システムでは，簡易なセンサを多数設けることで個々のセンサからのデータ欠損を許容しつつシステム全体の信頼性を維持する方針を取る．

さらに，ある程度事前に予測は可能であるが計画停電や機器の維持・管理など数日間にまたがる長期のデータ欠損についても検討を行わなければならない．

この問題に対して，予備実験として前述の 4 ヶ月分のデータ (151MB) を用いた検索において，ユーザが指定した検索期間に余剰区間を加えて検索を行った場合の追加コストを調査した．その結果，30 秒の余剰区間を設けた場合と 1 時間の余剰区間を設けた場合の時間差が 0.03 秒程度であり，1 時間程度であれば余剰区間の追加により検索時間の増加は許容可能であることが分かった．したがって，累積欠損時間が 1 時間未満であれば正しい OFFSET 検索は可能である．しかしながら，長期欠損に対しては，この手法単独での欠損対策は不完全である．

そこで，長期間のデータを保存するアーカイブデータに対しては，保存用のデータベース作成時にデータベースに格納された観測データを調査し，当該データの毎正時毎のオフセット位置を記録した

正時検索テーブルを別途作成し，観測データとともに保存することとした．検索時には，オフセットテーブルをまず検索し，検索開始 OFFSET 位置を正時起点とすることで，最大 1 時間の余剰区間付き OFFSET 検索と等価で，かつ，長期欠損にも対応可能な手法を考案し実装を行った．

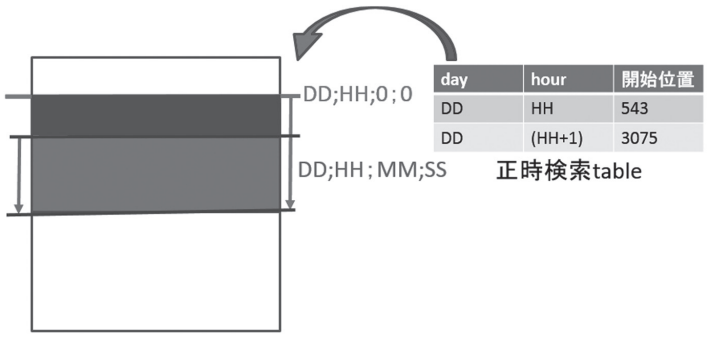

図-3 正時検索テーブルを併用したOffset検索

図 -3 に正時検索テーブルを併用した検索の概要を示す．この例では，ユーザが DD 日 HH 時 MM 分 SS 秒をオフセットとして指定した場合，正時検索テーブルから DD 日 HH 時 0 分 0 秒に相当するデータベース内のオフセット位置を求め，ユーザ指定のオフセット位置に MM 分 SS 秒（1 時間未満）の余剰検索区間を追加した OFFSET 検索を行う様子を示している．

3．3　短周期観測データの継続的アーカイブのためのデータ圧縮

観測データの長期アーカイブを可能にするためには，さまざまな観点でデータ圧縮法を検討しなければならない．例えば気象庁が公開する風速データは 10 分単位の平均風速，瞬間最大風速等が保存さており，我々の 1 秒毎の観測データを気象庁データと同様に 10 分単位に集計することで 1/600 に情報を圧縮可能であるが，これは非可逆圧縮であり圧縮後のデータから元のデータを復元することはできない．我々は，1 秒毎の短周期観測データに意義があると考えており，可逆な圧縮が必要であると考える．また，それぞれのデータの特性に応じた固有の圧縮アルゴリズムを開発すれば高い圧縮率を得ることも可能であるが，我々は単にデータをアーカイブするだけが目的ではなく，圧縮保存されたデータに対する Web ベースのリアルタイムなオンライン検索をも可能としたい．そのためには，単に圧縮率が高いだけでなく，データ活用時の解凍時間が短いことも圧縮法の検討において重要な評価指標となる．また，データをオープンデータとして公開し，第 3 者にデータの利活用を許可するためには，汎用的な圧縮・解凍ツールを利用したアーカイブが必要である．

そこで，そのような圧縮ツールの候補として Linux 環境下での汎用の圧縮ツール gzip，xz，ならびに bz2 を用いた圧縮の効果を調査した．評価に使用したデータは長期データ欠損を含まない 2018 年 1 月から 3 月のデータを用い，各月のデータと 3 ヶ月分のデータを 1 つのデータベースに統合したデータに対して，データの圧縮率，圧縮時間ならびに解凍時間を調査した．表 3 にその結果を示す．

いずれのデータに対しても圧縮率では xz が最高で gzip の約 4 倍，解凍時間については gzip の方が有利であり xz より約 2 倍高速である．bz2 は両者の中間的な位置づけである．圧縮時間については xz は最も不利であるが，圧縮はアーカイブ作成時に一度行うのみであり実用上問題とはならない．他の月についても同様の調査を行ったがほぼ同様の結果が得られている [3].

これらの結果から，圧縮ツールとしては xz を基本とすることに決定した．一方で，アクセス頻度が高いと思われる一定数のデータに

表-3 圧縮ツールの特性評価

期間	元データ		gzip	xz	bz2
2008 年 1 月	39.2MB	圧縮サイズ	15.2MB	3.36MB	10.0MB
		圧縮時間	3.93s	18.9s	5.05s
		解凍時間	0.39s	0.85s	1.36s
2008 年 2 月	35.7MB	圧縮サイズ	13.9MB	2.90MB	9.11MB
		圧縮時間	3.62s	16.9s	4.62s
		解凍時間	0.36s	1.05s	1.22s
2008 年 3 月	37.2MB	圧縮サイズ	15.1MB	3.91MB	11.1MB
		圧縮時間	3.58s	18.6s	4.70s
		解凍時間	0.38s	0.87s	1.33s
上記 3 ヶ月分	116.6MB	圧縮サイズ	44.3MB	11.3MB	30.4MB
		圧縮時間	56.3s	18.9s	4.70s
		解凍時間	1.16s	2.53s	4.36s

対しては gzip で圧縮したデータを，さらに，直近数ヶ月分は非圧縮のデータを併設することで検索システムの高速化を目指すこととした．将来的には，検索システムのアクセス統計に基づいて，gzip 保存のデータを動的に選択することを検討しているが，現在はいくつかの月を固定的に設定している．

3．3　アーカイブデータの公開と Web ベースの On Demand 検索システム

従来のシステムでは，クラウドシステム上に Web サーバを構築し，観測データの収集と検索を一元的に処理してきた．しかしながら，データベースへのアクセス集中や検索時間の観点から直近数ヶ月分のデータのみを検索対象としてきた．また，クラウドサーバの課金上の課題（通信量課金）から，長期アーカイブデータに対しては学内に別途設置した Web サーバを用いてデータの公開と検索を行うこととした（図 4 参照）．

観測データのリアルタイム収集と直近データの検索を行うサーバ（一次サーバと呼ぶ）では，データベース自体が 1 秒毎の短周期で更新されるため圧縮データの利用は困難である．また、随時データが更新されるため，前述の正時検索テーブルを用いた検索ができない．そこで，直近数ヶ月分のデータに対しては，非圧縮のデータを用いて Between 方式を用いた On Demand 検索システムを提供する．一次サーバのデータは，数ヶ月ごとの定期的な保守のタイミングで，バックアップするとともに初期化を行うことでデータベースの肥大化にともなう速度低下を回避する．

1 次サーバでバックアップされたデータは，データの長期アーカイブとアーカイブ検索を行う 2 次サーバに移し公開と検索の対象となる．2 次サーバでは，バックアップされたデータを 1 月単位のデータベースに分割し，正時検索テーブルを追加した後、xz 圧縮を行ってアーカイブする．この 1 月単位のデータは Web サーバから直接ダウンロードすることも可能である．このサイトでは，検索やダウンロードの参考のために，月単位でデータの概要をプロットしたグラフを数ヶ月分提示している．図 5 は原稿執筆時点で提示している 2020 年 5 月分の観測データ概要である．5 月 19 日付近で非常に強い風が吹いていたことがわかるが，過去の天気情報を調べると，5 月 19 日は急速に天気が悪化し雷雨が発生した日で

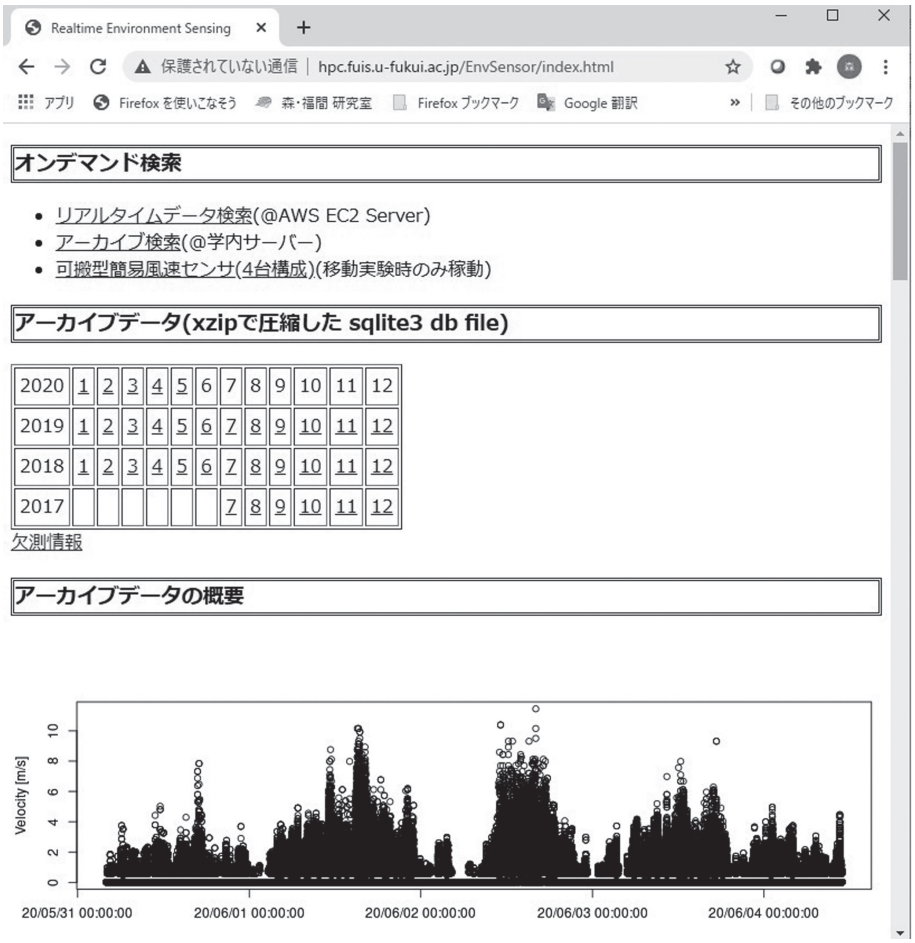

図4　WebベースOnDemand検索システム
(http://hpc.fuis.u-fukui.ac.jp/EnvSensor/index.html)

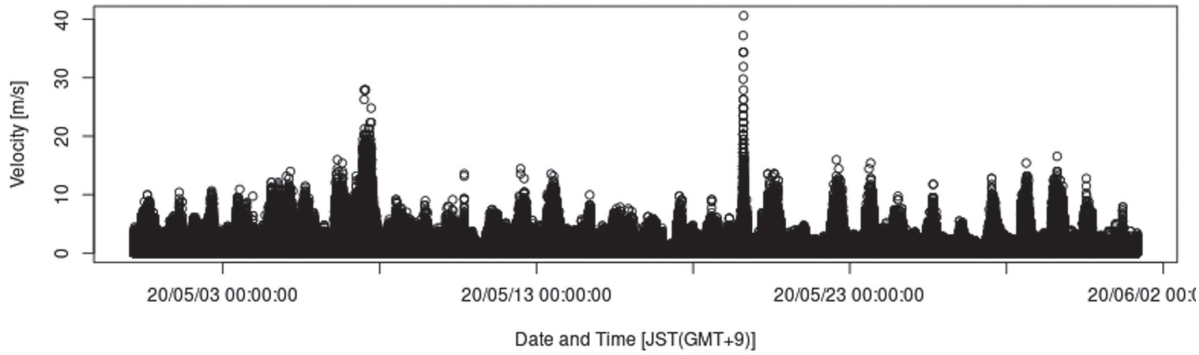

図-5　2020年5月の観測データ概要

あることが確認できた.

なお，2 次サーバでの月単位のデータ分割に際しては，AOE(Anywhere On the Earth) 時間を考慮した重複データ保存を行っている．具体的には世界中のどの国の Local Time で検索を行っても，同一月内のデータが 1 つの圧縮ファイルの解凍のみで検索可能となるよう，日本標準時と日付変更線との前後の時差を考慮して日本標準時での 1 月分のデータに前後あわせて 1 日分のデータを保存している.

検索期間が 2 カ月以上に跨ると複数のアーカイブデータにアクセスを行い，それぞれの検索結果を連結した後にブラウザ上に提示する．例えば 2 カ月に跨る検索では，最初の月のデータでは検索開始位置以降すべてのデータを出力し，翌月のデータに対しては前月分のデータに含まれる AOE 時間を考慮したデータとの重複部分を除いた位置から検索終了時間までの結果を出力する．3 カ月以上に跨る場合は，最初と最後の月以外の月ではデータの先頭から重複分（1 日分に相当）を除いたものを出力する．図 -6 に 3 カ月に跨る検索時のアーカイブ検索の様子を示す.

複数ファイルの解凍，検索，連結そしてブラウザへの提示の一連の動作をユーザが意識する必要はない．ただし，提示する情報量が膨大となるため，ユーザが検索に利用するブラウザの設定で決まるデータ量を超える検索結果の表示は不可能である．2020 年 2 月末の時点では 3 カ月に跨る実質 2 ヶ月分のデータ検索ができることを確認出来たが，同年 9 月上旬の原稿執筆時においては，実質データ量が一ヶ月分（概ね 50MB）を超えると検索結果の表示が中断する現象が発生している．原因は調査中であるが, COVID-19 の影響をうけたセキュリティ対策強化の影響が考えられる．なお将来的には，ブラウザへの直接情報提示ではなくユーザのローカルディスクへのファイルダウンロードを可能にするインタフェースを設けることでこの問題は回避することができると考える.

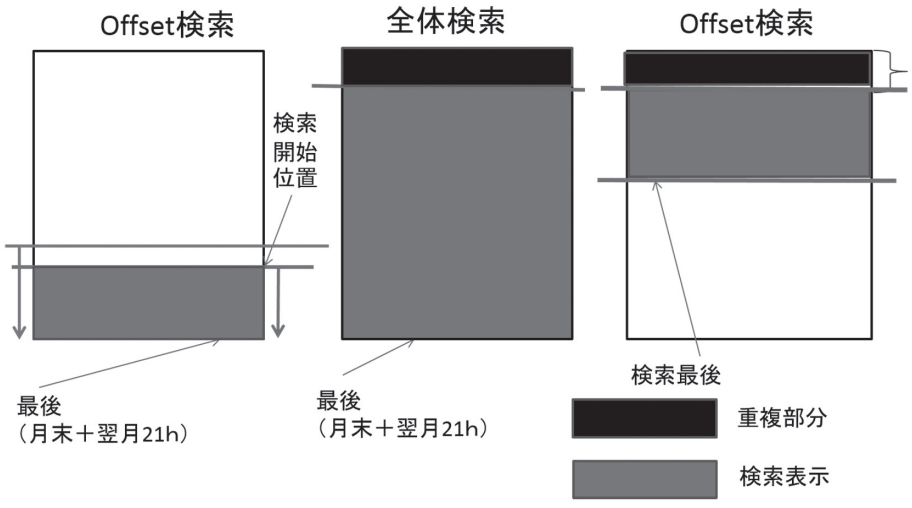

図-6　3カ月に跨る検索

4．Web ベースの On Demand 検索インタフェース

表-4 アーカイブ検索サーバ諸元

CPU	Intel Xeon W3680 3.33GHz (6 コア , 12MB Cache)
Memory	24GB
Disk	2TB
OS	CentOS 7.7.1908
http サーバ	Apatch 2.4.6
PHP	Ver.5.4.16 (Mem Limit 128MB)
DB	Sqlite3 3.6.20

検索インタフェースは検索対象日に応じて 2 つの検索サーバを使い分けることとした．毎秒観測データが更新される一次サーバ（Amazon EC2 クラウドサーバ　東京リージョン）には，前述の通り，直近数カ月分（保守等でサーバを停止した時点以降）のデータが格納されており，データベースへのアクセス競合によるデータ欠損を回避するため，検索負荷をなるべくかけない簡易な検索のみを提供する．具体的には，Between 方式を用いた直近 N 回分の観測データ表示を行うための検索インタフェースを設けるとともに，10 分毎に更新される過去 100 分の風況図を提示している．

長期アーカイブデータの検索に対しては，通信量課金の問題から学内に設置したアーカイブ検索サーバで検索サービスを提供する．2017 年 7 月の観測開始以降，直近の保守作業までのデータ検索が可能である．表 -4 にアーカイブ検索サーバ (2019 年度に故障による不要申請がだされたサーバを修理し再利用したものである) の仕様を示す．また図 -6 に検索サイトの様子を示す．

アーカイブ検索では，検索期間を自由に設定可能であり，現時点では検索開始点と期間（応答時間を考慮した選択方式）を指定した期間指定検索と検索開始点と終了点を指定した両端指定検索の検索が可能である．アーカイブ検索サーバでは，毎秒毎のデータ登録が発生しないため，検索時間を要する複数月にまたがる検索も可能である．現時点では単純な期間のみの指定であるが，複雑な検索条件を適用する検索インタフェースも今後検討する予定である．なお，検索期間の検討や長期欠損の有無を確認するための補助材料として，アーカイブデータの月単位の概要を把握可能なイメージデータ（図 -5 はその一例）の閲覧やアーカイブ期間中に発生したいくつかの特徴的な気象イベントを指定したイベント指定検索も可能であり順次拡充中である．

5．まとめ

1 秒間隔という高頻度でセンシングを行う風速センサからの観測データをクラウド上のサーバで情報集約し，長期にわたって継続的にアーカイブするとともに，アーカイブデータに対するオンデマンド検索を可能にするシステムの構築について報告を行った．現在はアーカイブデータの更新時に各種データの整合性検証や Web サイトの更新を人手で行っているが，今後はこれらの自動化についてもけんとうをおこなっていきたいと考える．また, 現在は本学内のセンサ情報のみを集約しているが, さまざまな場所からの多数のデータを集約することも考慮した検索インタフェースの改良を検討していきたいと考える．

謝　　辞

日頃様々な面でお世話になった情報・メディア工学専攻　森・福間研究室の皆さまに深く感謝いたします．なお, 本研究の一部は, 平成 28 年度ならびに令和元年度の地域環境研究センター研究費支援, ならびに科学研究費補助金（20K11744）の助成を受けて実施した．これらの助成に対してここに謝意を表します．

図-6　アーカイブ検索ページ
（http://hpc.fuis.u-fukui.ac.jp/EnvSensor/A19/single/index.html）

参考文献

1）ZHOU Rui, ZHANG Jiaochao, FUKUMA Shinji, MORI Shin-ichiro, 2017, “Case study: Development of a time-critical IoT system for environment sensing,” 平成 29 年度電気関係学会北陸支部連合大会予稿集.

2）風杯型風速計「WeatherSensorAssembly/n80422」
https://www.argentdata.com/files/80422_datasheet.pdf

3）浅越　陽一, 2020, 簡易無線センサ網を用いた短周期観測データの継続的アーカイブと Web ベース検索システムの高速化, 福井大学工学部情報・メディア工学科卒業論文