D-028

データ取得制限のある Deep Web からのサンプルデータ収集方式

杜翔 † Xiang Du　大森匡 † Tadashi Ohmori　藤田秀之 † Hideyuki Fujita　邱原 † Yuan Qiu　新谷隆彦 † Takahiko Shintani

1. 研究の背景と目的

Deep Web とは，インタネット上にある情報のうち，通常の検索エンジンが収集できない情報を持つ Web サイトを指す総称である．今までの Deep Web Crawling 手法の研究では，ニュース記事サーバなどの Textual Data Source を対象としていることが多い．一方，今日では Web 上のデータとして，生成時刻と位置情報を持つ Spatial Data が増えていて，Deep Web における Spatial Data の Crawling も大きな課題となっている．検索 API 制限と位置情報などの属性を加えた Spatial Data の Crawling は，単純に Textual Data の Crawling 手法のストラテジを適用できない．特に，単位時間あたりの API アクセス回数の制限が強い．本研究では，こうした空間 Web Object の共有サービスを対象に，外部アプリケーションが全データを取得するのではなく，サンプルを上手く集めることによって，効率よく空間情報抽出を行えるようなサンプル収集の技法を提案する．

2.Flickr 写真の 100% Crawling の紹介

Flickr は，大規模な写真共有サービスである．プライバシーの設定に従い，ユーザーがアップロードした写真には，位置とタグなどの情報が付けられる．画像そのものは，Textual Data とは違って，直接検索することはできず，タグや写真タイトルなどを Keyword として検索することが一般的である．そして，Flickr の写真データは Flickr の Backend-Server に保存されていて，Flickr が提供している API を用いてしかアクセスできない．本研究では，Deep Web の Spatial Data の Crawling 手法を対象にしているので，Flickr 膨大の写真データの中で，位置情報とタグが付いているデータだけを抽出して，以下の研究を行う．

まず Flickr 写真の Crawling システムについて説明する．図 1 に示した通り，Flickr Search API にクエリを送って，得られた写真データを Photo Database に保存する．API を提供している側による幾つかの API 制限がある．藤田は文献 [1] において，アクセス制限を克服するために，検索パラメタとして時間と空間の分割により，可能な限り API を最大限に利用して，全データを取る技法を提案している．そこて，本章では藤田の技法に基づいて 100%データを収集して，適切なサンプルデータ収集が可能なのかを分析する．

収集プロセスの API 制限について具体的に説明する：

1. ある API Key に対して，クエリを送って API にアクセスする回数は最大一時間 3600 クエリに制限されている．
2. あるクエリ検索に対し，いかなる結果の数があっても，user が指定しているソート順の Top 4000 個しか返さない．
3. Bounding Box 条件をつけていない検索クエリ，つまり検索キーワードや，時間条件だけを指定した検索クエリは Standard Photo Query と言う．こうした検索クエリはページごとに最大 500 個アイテムを取得することができるが，本手法で用いる空間地域の設定 (Bounding Box の設定を on にする) を加えたクエリ検索はページごとに 250 個に制限されている．

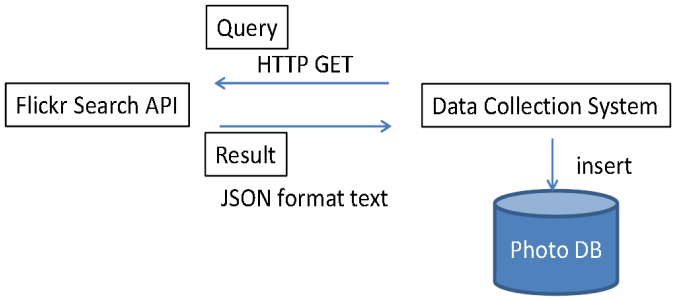

図 1: Data Collection System

上記強いデータ取得制限が存在する以上，どうやって Sampling で Flickr 写真をうまく収集するのかは本研究の課題として考えられる．

3.Quad-Tree による全件取得手法

Quad-Tree 構造は各内部ノードが 4 個の子ノードを持つ木構造であり，四角形の形を使って葉ノードまで分割していく．Quad-Tree の Region 分割の例を図 2 に示した．

各セルに Capacity の上限 (CELL_SIZE と記す) があり，容量の最大値に達すると，セルを 4 等分する．図のように，CELL_SIZE 4000 で空間領域を適当なセルに分割する．

3.1.STCrawling 収集法の説明

Quad-Tree 構造を用いた STCrawling 手法のアルゴリズムを説明する．

データ収集に用いる API の仕様として，検索条件に領域 S，期間 T，ページ番号 i を指定してアクセスすると，検索結果として，S, T に含まれる空間オブジェクトの総数 N, ページ総数 N_p, i ページ目の空間オブジェクト集合 A_i が得られる．ページとは，検索結果の集合を分割した部分集合を指す．1 つのページには，最大 τ 件 (API の仕様で定められる) の空間オブジェクトが含まれており，総数 N の空間オブジェクトが，N_p ページに分けて提供される．API へのアクセス一回につき，1 ページづつしか取得できない．したがって，全ての検索結果を取得するためには，ページ番号を変更しなが

†電気通信大学，UEC

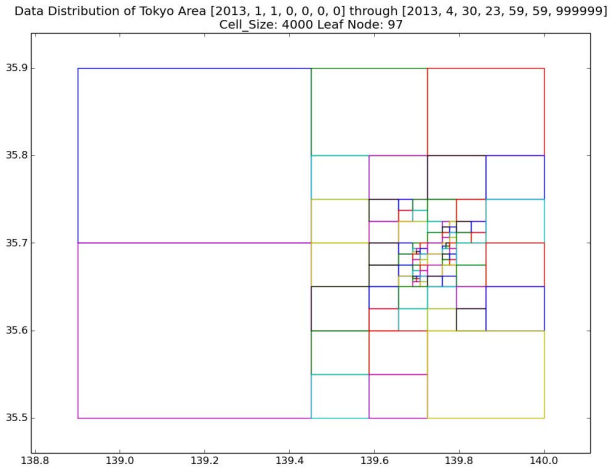

図 2: 東京都 4ヶ月間データの空間分割のみで Quad-Tree 構造の調査

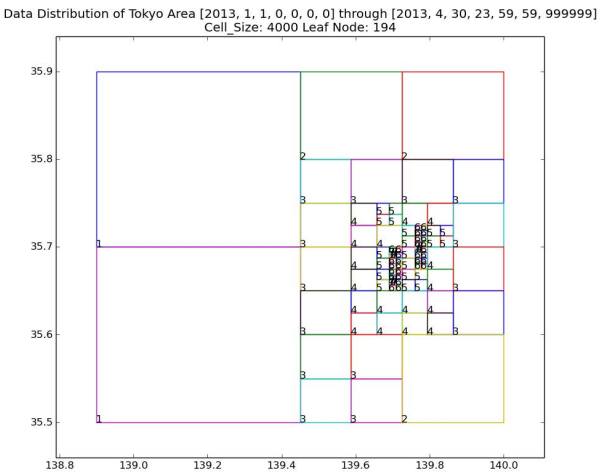

図 3: CELL_SIZE 4000，空間分割のみで 4ヶ月データの空間分割深さの調査

ら，N_p 回 API にアクセスする必要がある．また，指定領域と期間のオブジェクト総数 N が API 制限 4000 個を上回っている時，Quad-Tree 構造による領域の 4 等分と期間の 2 等分によって，各分割したセルで上記 Crawling 手順を行う．

3.2.STCrawling 手法による全件収集の実験

空間分割と時間分割両方を適用した STCrawling 手法を用いて，東京都 2013 年各時間帯のデータ収集実験を行った (表 1)．

4. 本研究 Sample 収集手法の提案

表 1 に示した通り，東京都 2013 年 1-4 月までの 4ヶ月間のデータは 7 万 5 千件あることが分かる．そのデータを 100%集めるには，API アクセス回数は 571 回，実行時間は 30 分ほどかかった．そこで，もっと API コール回数を削減するには，データを 100%取得するではなく，サンプルで収集したい．データ収集プロセスのアクセス回数の削減と取得データ品質を一定程度に保つことを目的として，Density-Based Sampling 手法を提案する．

分割深さ (Split Depth，SD と記す) はセルの分割回数を示している．図 3 のように，CELL_SIZE 4000，空間分割のみの条件で東京都四ヶ月間のデータを確認すると，深さ 7 まで分割していることが分かる．

許容誤差は，定めた大きさのセルの中に，タグ C が何個存在するのかに関わらず，C を持つデータが少なくとも一つ，Sample としてそのセルに存在すればよいということである．本研究で，許容誤差のセルは図 4 分割深さ $SD = 4$ のセルに決める．

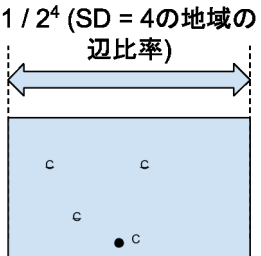

図 4: 分割深さ $SD = 4$ のセル

4.1.Density-Based Sample Crawling 手法の方針

上記 Sampling 手法の属性の説明に基づいて，本研究 Density-Based Sampling 手法の方針を説明する．

- 図 5 のように分割深さ 4 のセルを基づいて API コールを 2 回して，500/4000 つまり 12.5%の Sampling 基準で Crawling を行う．

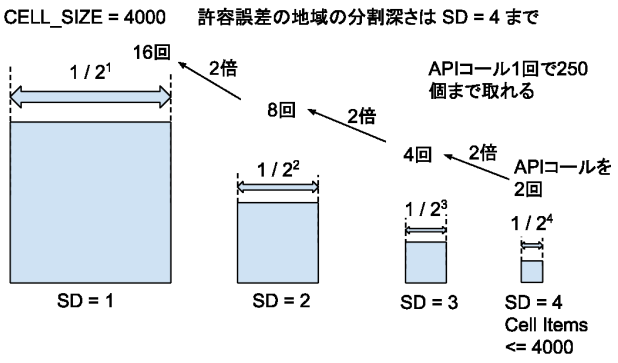

図 5: Density-Based Sampling 手法の許容誤差地域までの説明

- そして分割深さ $SD = 3$ のセル (辺比率 $\frac{1}{2^3}$) は，深さ 4 の 2 倍コールを使ってつまり 4 回まで，ページごとの Crawling を実行する．4 ページまでのセルは全件取ることとして，4 ページ以上ある場合

表 1: STCrawling 手法による Flickr Photos の Crawling 情報

データベース	対象期間	対象地域	実行時間	Access 回数	Count
01010131T	一ヶ月 (2013 年 1 月)	東京都	8 分	138	17013
01010228T	二ヶ月 (2013 年 1-2 月)	東京都	10 分	249	28479
01010430T	四ヶ月 (2013 年 1-4 月)	東京都	30 分	571	74649
01010630T	半年 (2013 年 1-6 月)	東京都	48 分	909	122332
01011231T	一年 (2013 年)	東京都	75 分	1687	254262

は，Sampling として残りのデータを捨てる

- このような基準を繰り返して，分割深さ $SD = 1$ のセル (辺比率 $\frac{1}{2^1}$) で，16 回まで，つまり 100%取ることとする
- 上記 Step で $SD = 4$ まで分割して，その Cell が 4000 個アイテムを越えている場合，$SD \geqslant 5$ で分割をせずに，時間分割を導入する．各分けられた期間で分割深さ $SD = 4$ までの Crawling 方針を適用して，少なくとも $\frac{16}{2^{SD-1}} = 2$ 回までの Crawling を実行する

4.2.Density-Based Sampling 手法の評価方式

検索地域の全域を許容誤差の地域の大きさまで平均的に Grid 分割する．本研究の場合，4 回の分割まで行うので，$2^4 \times 2^4 = 16 \times 16$ の Grid を用意する．そして，Grid ごとにタグを調べるが，タグの数に関わらず，存在することだけで正解 Grid だと見なす．Raw Data と Density-Based Sampling 手法のタグで正解 Grid の数 C を数えることによって，本 Crawling 手法の正解率 $P = \frac{\text{サンプル上で K がいるセル数}}{\text{Raw Data に K がいるセル数}}$ を測る．

5.Density-Based Sampling 手法の実験と評価

本研究で提案した Density-Based Sampling 手法を用いて，東京都 4ヶ月間の Sample Crawling を行った (表 2).

表 2: 東京都 4ヶ月の Density-Based Crawling 情報

実行時間	Access 回数	Count
6 分	209	27177

表によると，データ数で約 36%の Sampling になっていることが分かる．そして，上記評価方式によって，本手法と 40% Random Sampling の正解率の比較を図 6 に示す.

4ヶ月間データの中，タグが 250 以上出現している部分 (上位 300 位タグまでである) は Random Sampling より精度が上回っていることが分かる．そして，タグの出現回数が 250 以下に減っていきながら，二つの手法の精度が予測不可能な状態になる傾向がある．Crawling 上で，250 アイテムは 1 ページで取れるから，そのようなタグの検索は Density-Based Sampling を使わず，直接 API に 1 回アクセスするべきである．

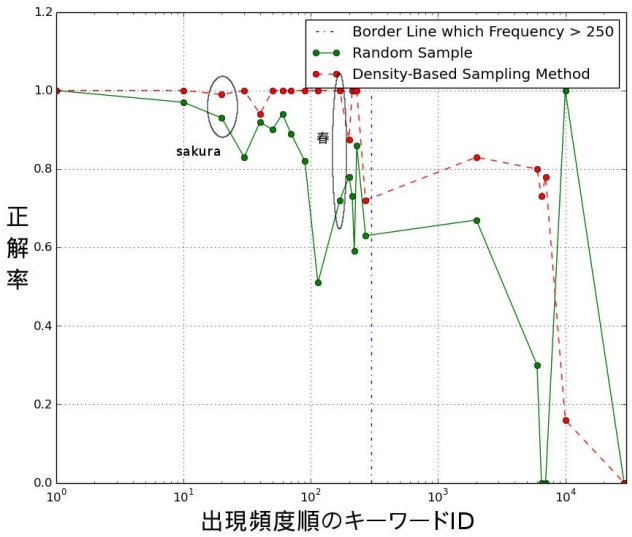

図 6: 4ヶ月データで Density-Based Sampling と Random 40% Sampling の Precision の考察

6. 結論

本研究は，Flickr 写真のデータを Spatial Data の例として，Spatial Data の Deep Web を対象として，適切なサンプルデータのみを収集する Crawling 手法を検討した．提案した Density-Based Sampling 手法は Crawling コストを減らし，一定の品質を保つことを目的としている．それぞれ違う密集性の地域で，それぞれ違う回数の Density-Based Sampling を行うことによって，全件収集より低いコスト，短い時間で Sample Data を抽出することが可能になる．

参考文献

[1] H. Fujita. Geo-tagged Twitter collection and visualization system. *Cartography and Geographic Information Science*, 40(3):183–191, June 2013.