1G3-OS-08b-4

大規模異分野データ横断検索における時空間スコア計算法

Spatio-Temporal Scoring Method for Large-Scale and Heterogeneous Data Search System

竹内 伸一 Shin'ichi Takeuchi　杉浦 孔明 Komei Sugiura　赤星 祐平 Yuhei Akahoshi　是津 耕司 Koji Zettsu

情報通信研究機構
National Institute of Information and Communications Technology

Publishing data to open access repositories has been a global trend in many fields such as governments and science. However it is difficult to use open data together from different domains because these open data is now managed individually and it is difficult to search datasets over domains. For solving this problem, we propose novel open data search system, named Cross-DB Search System. This system takes combination of the spatial, temporal, and text information as input query, and searches open data through several domains. To find more related open data, Cross-DB Search System uses query expansion named Spatial/Temporal/Text Pseudo Relevance Feedback (ST-PTRF). This method expands spatial, temporal, and text information of the search query using initial search result based on the similarity of the datasets. In this paper we discuss about the efficient definition of the similarity of the datasets and spatio-temporal scoring method based on the similarity.

1. はじめに

地質学や海洋学など自然科学の様々な分野において，調査によって得られた科学データを公開する，オープンデータ化が以前から取り組まれている．さらに近年，公的資金に基づく研究によってもたらされた科学データは公的に利用可能であるべきと見なされはじめており，OECD による利活用のガイドラインが発表されている [OECD 07]．一方で政府や行政機関が保有する公共データを公開し利用を促進する動きも近年広まっており，また経済においては，オープンデータ活用ビジネス市場が 1 兆円規模と試算されている [JETRO/IPA 13]．オープンデータを取り巻く環境はますます活発になっており．オープンデータの活用による更なるイノベーションが期待されている．

しかしオープンデータ検索を行う場合には主にデータ公開元が提供する検索システムが用いられており，Web 検索のようにオープンデータを網羅的に対象としたデータ検索の研究は近年ようやく取り組み始められた段階にある．膨大なオープンデータの中からいかに目的とするデータ，さらにはそれと関連する新たなデータをいかに見つけるかがオープンデータ活用のために取り組む必要のある問題のひとつである [Simmhan 07]．

オープンデータの中でも特に科学データ検索の際に問題となるのが，Web ページなどテキスト情報を対象とする検索システムにおいて用いられる自然言語処理技術のみでは十分な検索が行えない点である．科学データが持つテキスト情報量は限られており，テキスト情報を対象とする一般的なキーワード検索では結果を充分に発見することが困難である．一方で科学データはデータを観測した際の時間情報や空間情報など，Web ページにはない種類のメタ情報を持つという特徴がある．この問題を解決するため，筆者らはオープン科学データを対象とした分野横断データ検索手法 [Gonzales 13] を提案している．

本稿ではオープン科学データ検索の性能向上のための効果的な時空間情報間のスコア計算法の提案を行う．時空間スコア計算の元となる時空間情報間の距離に関して，時空間検索に適したデータセット間の距離定義について検討する．また，それらの違いが検索性能に及ぼす影響を検索結果の Precision/Recall を用いての評価を行う．本稿での提案手法によって以下の項目を実現する．また，本研究で開発したオープン科学データ検索システム Cross-DB Search System は http://dataeyez.org/crossdb/ で利用可能である．

- 科学データ検索のための時空間クエリ拡張手法．時空間条件を時空間情報をクエリの時空間的な意図とみなし，検索範囲の絞り込むために用いるのではなくテキストクエリと同様に扱う．これにより擬似適合性フィードバックによるクエリ拡張 [Buckley 93] のようなテキスト対象の技術を時空間情報へ応用することが可能となる (2. 節参照)．
- 検索性能向上のためのデータセットの時空間スコア計算法．(3. 節参照) データセットの時間/空間での分散を考慮したデータセット間距離を用いて検索結果のスコアリングを行う．評価実験により，適切なデータセット間の時空間距離を用いることで検索性能が向上することを示した (4. 節参照)．

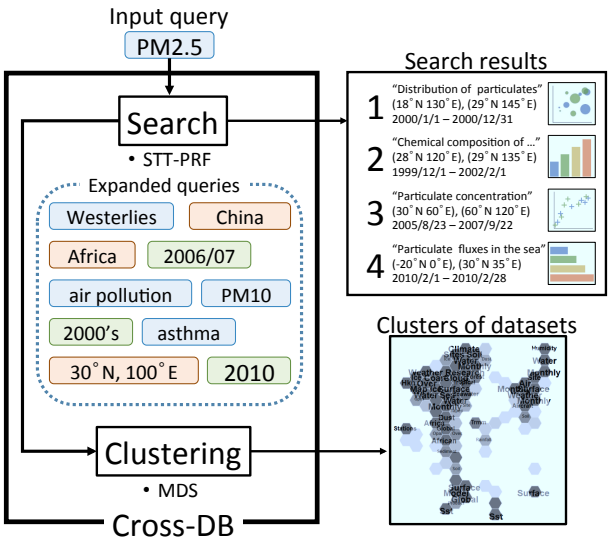

図 1: Cross-DB Search System の概要

連絡先: 連絡先：竹内伸一，情報通信研究機構，〒 619-0289 京都府相楽郡精華町光台 3-5，s.takeuchi@nict.go.jp

2. Cross-DB Search System

本節ではオープン科学データを対象としたデータ検索システム Cross-DB Search System の構成とクエリ拡張法について述べる．図 1 に Cross-DB Search System の概略を示す．Cross-DB Search System の特長は (1) 時空間情報に基づく擬似適合性フィードバックを用いたクエリ拡張技術と，(2) 検索結果のクラスタリングである．Cross-DB Search System は現在約 80 万件のオープン科学データを検索対象とし，一般的なキーワードだけでなく時空間情報をクエリとして検索を行うことが可能である．また，検索の結果として得られたデータセット集合に対し，時間/空間/オントロジー的な距離でクラスタリング，可視化を行うことで関連性のあるデータセットの発見を補助する．

2.1 システム概要

図 2 に Cross-DB Search System の Web GUI を示す．

画面左側の検索結果可視化部では，検索によって得られたデータセット集合を二次元上のセルに配置していく．このとき各データセット間の時間/空間/オントロジー距離に基づいて，距離が近いデータセットは近くのセルに，遠いデータセットは遠くのセルに配置される．具体的には時間/空間/オントロジー距離の三次元で定義されたデータセット間距離を用い，多次元尺度構成法によって得られた二次元上の座標に基づいて所属セルを決定していく．

画面右側では検索によって得られたデータセットの時間/空間/オントロジー上の分布を表示する．データセットの時間分布は時系列上に、空間分布は地図上にそれぞれ表示される．オントロジー上の分布に関してはデータセットと科学用語オントロジーである SWEET Ontology[*1] 中の関連するコンセプトとの関連付けを行い，各コンセプトをノードとして構成されるグラフ上に対応するデータセットを表示する．

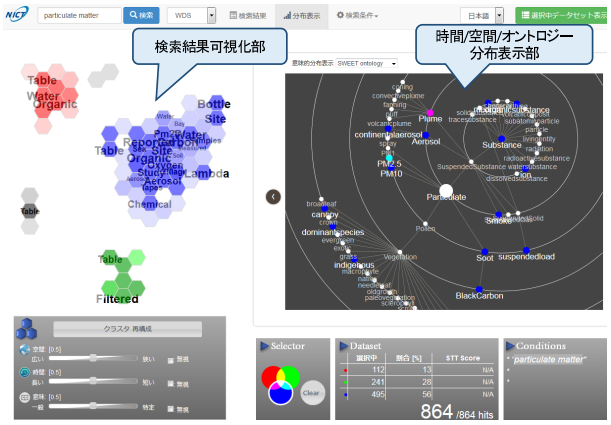

図 2: Cross-DB Search System の Web GUI

2.2 時空間情報を用いた検索クエリ拡張

Cross-DB Search System は検索対象である科学データの，テキスト情報は少ないが観測データの時空間情報を持つという特徴を踏まえ，時空間情報を用いた検索クエリ拡張を行う．本節では時空間情報を併用した疑似適合性フィードバック (Spatio-Temporal and Text Pseudo Relevance Feedback, STT-PRF) について述べる．

1. ユーザーからの入力キーワードをテキストクエリとして検索用インデックスに対し一度目の検索を行う．テキストスコア計算部は各データセットのテキストクエリに対するスコア ϕ_k の計算を行い，スコア上位のデータセットから順に結果として検索部に返される．本稿では ϕ_k としてデータセットとテキストクエリの TF-IDF ベクトル間のコサイン距離を用いた．

2. 次にクエリ拡張を行う．検索結果の上位 L 件をテキストクエリに適合していると仮定し，仮適合データセット集合 Y_L を決定する．テキストクエリ構築部は，Y_L に含まれるデータセットのテキスト情報を追加のテキストクエリとしてクエリの再構築を行う．STT-PRF はさらに Y_L から構築される空間クエリとして Y_L 中の各データセットの空間情報の集合を用い，同様に時間情報の集合を時間クエリとする．これらを統合して空間/時間/テキストクエリ (STT クエリ) を構築し，これに基づき二度目の検索を行う．

3. 二度目の検索ではインデックス内の各データセットに対し，STT クエリに対するスコアが計算される．仮適合データセット集合 Y_L から作成した空間/時間クエリに対するデータセット y の空間スコア $\phi_s(y)$ および時間スコア $\phi_t(y)$ を計算し，$\phi_k(y)$ と統合することでデータセットの統合スコアを決定する．最終的に統合スコア上位のデータセットから順に結果としてユーザーに提示される．時空間スコア計算法に関しては 3.2 節において述べる．

3. データセット間の時空間スコア計算法

3.1 科学データの時空間情報

データセットによって表されるデータそのものとは別に，データセットそれ自体の情報はメタデータとして扱われる．科学データが持つメタデータの例としてはデータセットの題目，作成者，概要，観測対象項目といったテキスト情報の他に，データを観測した位置や時刻などで表される時空間などが含まれる．特に時空間情報は科学データセット検索において有益な情報となる．本節ではデータセットが持つ時空間情報の特徴について述べる．

データセットの時空間情報は一般的に始点 $\boldsymbol{x}_b$ と終点 $\boldsymbol{x}_e$ とで決定される範囲として表現されるものとする．このとき，時間情報は一次元の時系列上の範囲として，空間情報は二次元平面上の範囲として表される．

時間情報に関してはそれぞれ観測開始時刻および終了時刻が相当し，例えば 1990 年 1 月 1 日から 2000 年 12 月 31 日にわたって観測されたデータで構成されるデータセットであれば $(\boldsymbol{x}_b, \boldsymbol{x}_e) = (1990/1/1, 2000/12/31)$ となる．空間情報に関しては，緯度と経度それぞれに対しデータの観測範囲から始点と終点を定める．緯度に関しては南端，北端をそれぞれ始終点とし，経度に関しては西端，東端をそれぞれ始終点とする．データセットの空間情報は緯度及び経度の範囲で示される二次元上の領域として表現され，南西端及び北東端を始終点とみなす．

3.2 時空間情報の間の距離定義

データセットが持つ時空間情報の間の距離定義に関してはこれまでも研究が行われているが [Wang 06]，科学データ検索において特に重要となるのは範囲をもった時空間情報の距離をいかに適切に表現するかである．本節ではこれまでに提案した距離定義 [Takeuchi 14] を拡張し，複数の分布および距離定義を用いた場合のデータセット間距離について検討を行う．

*1 http://sweet.jpl.nasa.gov/ontology/

前節で述べたように，データセットの時空間情報は始点 $\boldsymbol{x}_b$ および終点 $\boldsymbol{x}_e$ の対で与えられる．ここでデータセットの時空間情報を一様分布で表現することを考える．$\boldsymbol{x}_b$ および $\boldsymbol{x}_e$ をもつ時間/空間情報を表現する一様分布の平均 $\boldsymbol{\mu}$ および分散 Σ は式 (1) で与えられる．

$$\boldsymbol{\mu} = \frac{1}{2}(\boldsymbol{x}_e + \boldsymbol{x}_b), \quad \Sigma = \frac{1}{12}(\boldsymbol{x}_e - \boldsymbol{x}_b)^2. \tag{1}$$

このとき時間情報は一次元の，空間情報は二次元の分布として表現される．データセット間の時間/空間情報の距離を，それぞれを表現する分布間の Bhattacharyya 距離，もしくは L_2 距離で定義する．二つの連続分布 p および p' 間の Bhattacharyya 距離 d_B および L_2 距離 d_{L_2} は式 (2) および (3) で示される．

$$d_B(p, p') = -\ln\left(\int \sqrt{p(x)p'(x)}dx\right), \tag{2}$$

$$d_{L_2}(p, p') = \int \left(p(x) - p'(x)\right)^2 dx. \tag{3}$$

データセットを表現する一様分布を同じ $\boldsymbol{\mu}$ および Σ で表される正規分布で近似する場合，正規分布 p_i および p_j 間の Bhattacharyya 距離は式 (4) のように表される．

$$d_B(p_i, p_j) = \frac{1}{8}(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)^\top \left[\frac{1}{2}(\Sigma_i + \Sigma_j)\right]^{-1} (\boldsymbol{\mu}_i - \boldsymbol{\mu}_j) + \frac{1}{2}\ln\left\{\frac{\det(\frac{1}{2}(\Sigma_i + \Sigma_j))}{\sqrt{\det(\Sigma_i)\det(\Sigma_j)}}\right\}. \tag{4}$$

時間/空間情報間の距離を定義することで，テキストの場合と同様に時空間クエリとインデックス中の検索対象とのスコア計算や結果のランキング等が可能となる．Cross-DB Search System におけるデータセット間の時空間スコアとして式 (5) を用いる．検索クエリの時間/空間情報を表す p_q を用いて検索インデックス中の科学データの時間/空間情報を表す p_t の時空間スコア y は

$$y = \exp\left\{-d(p_q, p_t)^2\right\}, \tag{5}$$

と計算される．ここで d は式 (2)，式 (3) が示す距離定義 d_B, d_{L_2} のいずれかを表す．

4. 評価実験

本節ではデータセットの時空間情報を表現する分布及び分布間距離を変えた場合の，検索性能の変化について述べる．自然科学分野のデータセットを対象として 50 個の科学関連キーワードで検索し，分布/距離定義が検索結果に及ぼす影響を Recall/Precision を用いて評価する．

4.1 実験条件

評価実験において用いる検索用キーワードとして，Google Trends[*2] で公開されている実際の検索クエリおよび Cross-DB Search System の検索履歴の中から科学分野に関するキーワードを選択した．さらに最近の研究動向由来のキーワードとして Microsoft Academic Search[*3] の環境科学分野のキーワード，科学分野のオントロジーとして SWEET Ontology 内のコンセプト名から補完的に追加を行った．表 1 に実験で用いた検索用キーワードを示す．

*2 http://www.google.com/trends/
*3 http://academic.research.microsoft.com/

表 1: 評価実験用検索キーワード

high temperature	atmospheric circulation	air quality
marine biology	climate variability	boundary current
sediment	interannual variability	global climate
water cycle	sea level pressure	natural gas
sedimentary rock	sea surface temperature	ocean circulation
climate change	water quality	ocean current
southern oscillation	carbon cycle	precipitation
ice sheet	particulate matter	black carbon
acid rain	coastal waters	loop current
aerosol	ozone	tsunami
desert	heavy metal	hurricane
global warming	environmental impact	trade wind
greenhouse gas	water pollution	ozone hole
pollution	soil pH	ash flow
air pollution	acid deposition	tidal wave
glacier	boreal forest	typhoon
deforestation	species richness	

評価実験の検索対象データとして地球科学に関する科学データレポジトリである Pangaea [*4] が公開している科学データセットを用いた．Pangaea のデータセット検索システムを用いて表 1 内の各キーワードで検索を行い，得られた検索結果の上位 120 件をそれぞれのキーワードに対する検索対象のデータセット集合とした．各データセットには環境科学分野の修士号を持つ作業者 3 名によって 4 段階の関連度が付与されている．関連度はキーワードとデータセットが全く関連しない場合は 0 が，非常に関連する場合は 3 が，その中間の場合に 1 または 2 があてられる．さらに関連度が 2 または 3 のデータセットをキーワードに関連すると見なし，0 または 1 のデータセットをキーワードに関連しないと見なした．

本実験において検索対象となる科学データセットが持つテキスト情報としては題目，著者名および概要がある．これらのなかで概要は自然言語処理を効果的に用いるのに十分なテキスト量を持つため，キーワードで検索を行う場合に重要となる．表 2 に Pangaea がもつデータセットのメタデータに各種の情報が含まれている割合を示す．約 1.7% のデータセットのみが概要を持っており，これはキーワードのみでは十分な検索が行えないことを示唆している．一方で時間情報は約 73.3% のデータセットが，空間情報に関してはほぼ全てのデータセットが持っており，これらを検索に活用することで検索性能を向上させることが可能となる．

*4 http://www.pangaea.de/

表 2: 科学データセットが各種情報をメタデータに含む割合

データセット	個数	割合
全体	405,456	
概要あり	7,028	$R_a = 0.017$
時間情報あり	297,478	$R_t = 0.733$
空間情報あり	404,145	$R_s = 0.996$
時空間情報あり	297,037	$R_{st} = 0.732$

検討項目である時空間スコアの性能を評価するため，実験に用いるキーワード毎のデータセット集合に対し，概要を持つデータセットの割合 R_a を変化させることで，テキスト量が変化した場合の性能を評価する．全てのデータセットが概要を持つ場合を $R_a = 1$ とし，本実験では $R_a = 0.01, 0.02, 0.05, 0.10, 0.20, 0.50, 1.00$ を用いた．

4.2 実験結果

検索性能の評価指標として，検索結果の上位 n 件に基づいて式 (6) で定められる Precision@n (P@n) と Recall@n (R@n)

を用いる．

$$\mathrm{P@}n = \frac{\mathrm{tp@}n}{\mathrm{tp@}n + \mathrm{fp@}n}, \quad \mathrm{R@}n = \frac{\mathrm{tp@}n}{\mathrm{tp@}n + \mathrm{fn@ALL}}, \quad (6)$$

tp@n および fp@n はそれぞれ上位 n 件に含まれるキーワードに関連するデータセット数，関連しないデータセット数を表す．本実験では検索対象となる総データ数が判明しており，fn@ALL は検索に失敗した関連するデータセット数を表す．一般的な検索システムの評価では $n = 10$ が用いられることが多いが，科学データ検索においては広範囲な検索結果がもたらす多様性がより重視される．このため，本稿では $n = 30$ を用いた．

性能比較用のデータセット間距離尺度として，二つのデータセットそれぞれの時空間分布の中心間のユークリッド距離を用いた．これは分布の平均の L_2 距離に相当し，分布の分散は考慮していない．

各 R_a についてキーワード検索を行った際の P@30 及び R@30 の 50 キーワードでの平均値を表 3 に示す．

表 3: 分布/距離を変化させた場合の検索性能評価 (一部)

R_a	近似分布	距離尺度	P@30	R@30
0.01	正規	Bhattacharyya	**0.328**	**0.219**
	正規	L_2	0.321	0.206
	一様	Bhattacharyya	0.315	0.201
	一様	L_2	0.324	0.207
	平均	L_2	0.317	0.203
0.05	正規	Bhattacharyya	**0.332**	**0.261**
	正規	L_2	0.324	0.248
	一様	Bhattacharyya	0.322	0.246
	一様	L_2	0.322	0.236
	平均	L_2	0.322	0.243
0.20	正規	Bhattacharyya	**0.356**	**0.328**
	正規	L_2	0.346	0.306
	一様	Bhattacharyya	0.343	0.293
	一様	L_2	0.344	0.293
	平均	L_2	0.341	0.288

この結果から全ての性能指標において正規分布で表現された時空間情報間の Bhattacharyya 距離を用いた場合が最良となることが示された．本稿では一部の R_a のみの結果を示すが，全ての R_a について同様の結果が得られた．空間情報に関しては二次元の平面上にあるものとして (二次元の正規分布として) 扱ったが，実際には三次元球面上に表現される．このため特に極部分に位置するデータに関しては実際よりも広い領域として扱われてしまう問題がある．

図 3 に正規分布および Bhattacharyya 距離を用いた場合の R_a 毎の P@30 および R@30 をそれぞれ示す．図中の baseline はテキスト情報のみを用いてクエリ拡張を行わず検索を行った場合の結果を示す．実験結果から，時空間情報を用いたクエリ拡張によって，全ての条件において検索性能が向上したことが示された．

5. まとめ

本稿では時空間情報を検索クエリとして用いる分野横断的データ検索システム Cross-DB Search System のための，時空間情報間に基づくデータセット間距離定義の比較を行った．データセット間距離は検索クエリとインデックス中の各データセットから求められる時空間スコアに影響を及ぼし，検索結果に反映されるため，適切な距離定義を用いることで検索性能が向上する．評価実験によって，正規分布で表現された時空間情報間の Bhattacharyya 距離を用いることで Precision 及び Recall で評価される検索性能が向上することを示した．

今後の課題としては，4.2 節で述べた極付近での空間情報の歪みの解決，検索結果の時空間クラスタリングによる効率的なデータセット発見支援，科学データ以外のオープンデータとの分野横断検索時における粒度問題の解決などがある．

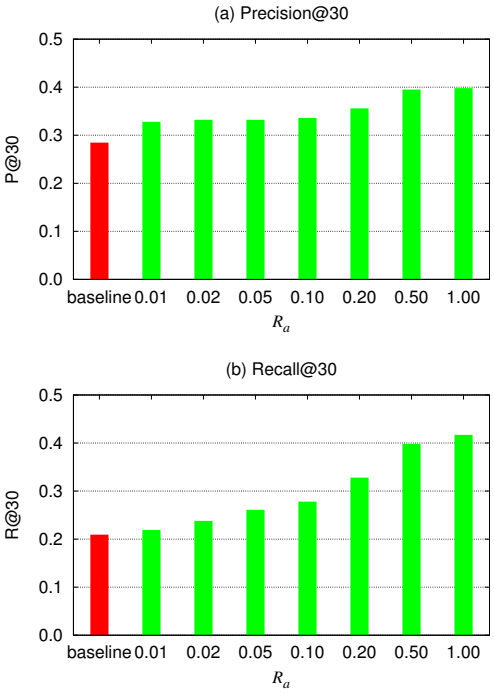

図 3: R_a を変化させた場合の (a) P@30 および (b)R@30

参考文献

[Buckley 93] Buckley, C., Salton, G., and Allan, J.: Automatic retrieval with locality information using SMART, in *Proc. of the 1st Text REtrieval Conference (TREC-1)*, pp. 59–72 (1993)

[Gonzales 13] Gonzales, E., Ong, B. T., and Zettsu, K.: Searching inter-disciplinary scientific big data based on latent correlation analysis, in *Proc. of Workshop on Big Data and Society (in conjunction with IEEE BigData 2013)*, pp. 9–12 (2013)

[JETRO/IPA 13] JETRO/IPA, : 米国オープンデータの動向 (2013)

[OECD 07] OECD, : *OECD Principles and Guidelines for Access to Research Data from Public Funding* (2007)

[Simmhan 07] Simmhan, Y. L., Pallickara, S. L., Vijayakumar, N. N., and Plale, B.: Data Management in Dynamic Environment-driven Computational Science, in *Proc. of the International Federation for Information Processing(IFIP)*, Vol. 239, pp. 317–333 (2007)

[Takeuchi 14] Takeuchi, S., Akahoshi, Y., Ong, B. T., Sugiura, K., and Zettsu, K.: Spatio-Temporal Pseudo Relevance Feedback for Large-Scale and Heterogeneous Scientific Repositories, in *Proc. of IEEE International Congress on Big Data*, pp. 669–676 (2014)

[Wang 06] Wang, S.-L., Xu, J., and Zeng, Q.: Using Statistical Similarity to Identify Corresponding Attributes between Heterogeneous Spatial Databases, in *Proc. of IEEE Asia-Pacific Conference on Services Computing*, pp. 194–199 (2006)