2E5-OS-25b-2

大規模異分野データ横断検索における 時空間情報を用いた疑似適合性フィードバック

Spatio-Temporal Pseudo Relevance Feedback for Large-Scale and Heterogeneous Scientific Data Search System

竹内 伸一[*1] Shin'ichi Takeuchi　赤星 祐平[*1] Yuhei Akahoshi　Bun Theang Ong[*1] Bun Theang Ong　杉浦 孔明[*1] Komei Sugiura　是津 耕司[*1] Koji Zettsu

[*1]独立行政法人 情報通信研究機構
National Institute of Information and Communications Technology

As larger and larger amounts of data are harvested, finding just the right piece of information out of this noisy and heterogeneous ocean of data remains challenging. Many widely adopted scientific data search engines continue to be mainly based on text semantics. However, it is not uncommon in scientific big data applications to face collected data that do not possess text information. In this scenario, search engines fail to retrieve potentially relevant data. In this work, we propose a novel pseudo relevance feedback method based on spatio-temporal and text (STT) information for scientific big data: STT-PRF. Although STT-PRF may simultaneously use STT information, we show that the missing values in space, time or/and the text are handled efficiently. We tested our STT-PRF method using the Pangaea repository. Experimental evaluations show that STT-PRF outperforms the standard baseline methods.

1. はじめに

Jim Gray によって提唱された「第四のパラダイム」は今日の科学研究手法のありかたについて，実験科学，理論科学，計算科学を経てデータ中心科学に到ったとしている [1]．第三のパラダイムにおいて個別に収集され分析されていた様々な科学データを集約し，それらを分析することで新たな科学的知見を発見するという考えは，増大する今日の計算資源によって実現が可能となった新しい科学的探求として今後発展していくと考えられる．一方で観測された科学データの共有や再利用，そのためのメタデータ定義，さらにはデータの蓄積，参照方法など解決すべき課題もあり，そのために国際的，学際的な連携が必要である．

科学データを再利用する際に最低限必要となるのが，それらのメタデータを管理するインデックスであり，またそれに基づいて科学者が必要とするデータを検索するシステムである．World Data System (WDS) [*1] や Pangaea [*2] は地質学や海洋学など地球科学の分野に特化した科学データ検索のポータルを提供している．

Web ページや文書データと大きく異なる科学データの特徴として，テキスト情報が豊富でない点がある．数値や画像がデータ本体であり，テキスト情報は題目や作成者名，ある程度の概要文などに限定される．このためテキスト情報に基づく従来の検索システムでは，本来なら入力されたキーワードに関係するデータを発見できない可能性がある．一方でメタデータからは収録された際の時空間情報や対象とするパラメータなど，テキスト以外の情報も多く含まれており，これらを活用することによってテキスト情報の不足を補うことが可能であると考えられる．

適合性フィードバック (Relevance Feedback, RF)[2] は検索システムの精度向上手法のひとつで，ユーザーは提示された検索結果のなかから入力したキーワードと関連するものを選択する．人手によって得られたキーワードと検索結果の対応情報を次回以降の検索に反映させることで精度の向上を図る．疑似適合性フィードバック (Pseudo Relevance Feedback, PRF) [3, 4] は RF の人手による部分を省略し，検索結果上位を仮に適合している文書とみなして入力クエリと併用する，クエリ拡張手法のひとつである．PRF の拡張については様々な研究がなされており，テキスト情報の活用としては協調的タグ付けシステムによるセマンティックアノテーションを用いたクエリ拡張 [5] がある．また，マイクロブログのテキスト情報を活用した動的な PRF[6, 7] では，時間情報の有用性が示されている．

筆者らはこれまでに，社会学と自然科学など全く異なる分野間のデータを組み合わせることで新たな科学的知見を得ることを目的とした，異分野データ横断検索システム Cross-DB Search System [8] を開発してきた．本稿では Cross-DB Search System の検索性能向上のため，テキスト情報に加えて時空間情報を併用した科学データ検索のためのクエリ拡張手法を提案する．本稿は以下のように構成されている．第 2. 節で提案手法である時空間情報を併用する疑似適合性フィードバックと，そのためのデータセット間時空間距離の定義について述べる．第 3. 節で性能評価実験と空間クエリ拡張例について述べ，第 4. 節でまとめと今後の課題について述べる．

2. 時空間情報を用いた疑似適合性フィードバック

本節では提案手法である時空間情報を用い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．STT-PRF はテキスト情報のみを用いる PRF に対し，時空間情報を活用するよう拡張したものである．

2.1 STT-PRF を用いた科学データ検索システム

図 1 に STT-PRF を用いた検索システムの概略図を示す．検索部はユーザーからのキーワード入力を受け取り，テキストクエリとしてインデックスから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

連絡先: 〒 619-0289 京都府相楽郡精華町光台 3-5 (独) 情報通信研究機構 ユニバーサルコミュニケーション研究所

*1 http://www.icsu-wds.org/

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

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．

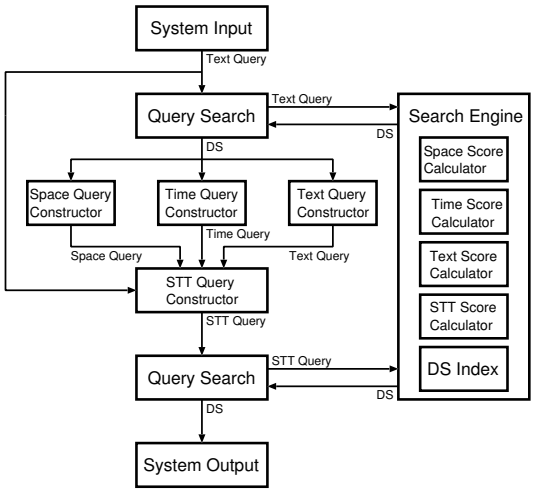

図 1: STT-PRF の概略図

次にクエリ拡張を行う．通常の PRF と同様に検索結果の上位 L 件を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．PRF と共通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，それに基づいて二度目の検索を行う．

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，空間/時間/テキスト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 $\phi_t(y)$ はそれぞれ式 (1) および (2) で与えられる．

$$\phi_s(y) = \exp\{-(\min_{y' \in Y_L} d_s(y, y'))^2\}, \tag{1}$$

$$\phi_t(y) = \exp\{-(\min_{y' \in Y_L} d_t(y, y'))^2\}. \tag{2}$$

ここで y' は Y_L に含まれるデータセットを表し，d_s および d_t は 2.2 節で述べる二つの空間/時間情報の間の距離を表す．データセット y の時空間およびテキストスコアから得られる統合スコア $\phi(y)$ を式 (3) に示す．

$$\phi(y) = w_s\phi_s(y) + w_t\phi_t(y) + \phi_k(y). \tag{3}$$

ここで w_s および w_t は空間情報および時間情報に対する重みを表す．二度目の検索ではインデックス中の各データセットに対し統合スコアの計算を行い，統合スコア上位のデータセットからから順に結果として検索部に返す．

2.2 データセット間の時空間距離

PRF は二度目の検索でもテキストクエリのみを用いるが，STT-PRF では空間/時間クエリを併用する．各クエリとインデックス中の各データセット間のスコアはそれぞれ式 (1) および (2) で与えられるが，これに際しスコア計算対象データセット y とクエリ中のデータセット y' との間の空間/時間情報に関する距離を定める必要がある．本節ではデータセットの時空間情報に基づく距離について述べる．

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．時間情報に関してはそれぞれ開始時刻および終了時刻が相当し，例えば 1990 年から 2000 年に得られたデータで構成されるデータセットであれば $(\boldsymbol{x}_b, \boldsymbol{x}_e) = (1990, 2000)$ となる．空間情報に関しては，緯度について南端および北端が，経度について西端および東端が相当する．

上記で定めた始終点に基づいてデータセットの時空間情報を正規分布で近似する．このときその平均 $\boldsymbol{\mu}$ および分散 Σ は，始終点 $\boldsymbol{x}_b$ および $\boldsymbol{x}_e$ から得られる一様分布の平均および分散に基づき

$$\boldsymbol{\mu} = \frac{1}{2}(\boldsymbol{x}_e + \boldsymbol{x}_b), \tag{4}$$

$$\Sigma = \frac{1}{12}(\boldsymbol{x}_e - \boldsymbol{x}_b)^2. \tag{5}$$

と定める．時間情報は 1 次元の，空間情報は 2 次元の正規分布として表現される．データセット y_i と y_j 間の空間/時間距離を正規分布間のバタチャリヤ距離 [9] と定める．二つの連続分布 p および q 間のバタチャリヤ距離は式 (6) で示され

$$d(p, q) = -\ln\left(\int \sqrt{p(x)q(x)}dx\right), \tag{6}$$

正規分布 y_i および y_j を用いた場合は式 (7) のように変形される．

$$d(y_i, y_j) = \frac{1}{8}(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)^\top \left[\frac{1}{2}(\Sigma_i + \Sigma_j)\right]^{-1}(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j) + \frac{1}{2}\ln\left\{\frac{\det(\frac{1}{2}(\Sigma_i + \Sigma_j))}{\sqrt{\det(\Sigma_i)\det(\Sigma_j)}}\right\}. \tag{7}$$

一般的に検索システムへの入力としての時空間情報は結果の絞り込みに用いられ，その始終点内に含まれる結果のみを返すことが多い．データセット間の距離を定義することで，クエリに対する距離計算が可能となり結果のランキング等が実現できる．

3. 評価実験

本節では提案手法である STT-PRF の性能評価実験について述べる．自然科学分野のデータセットを検索対象として複数のキーワードで検索を行い，検索結果から求めた評価指標で STT-PRF の有効性を示す．

3.1 実験条件

表 1 に評価実験で用いた 20 個の検索キーワードを示す．検索キーワードは自然科学に関するもので，科学分野のオントロジー，最近の研究分野，実際の検索キーワードの 3 種からそれぞれ収集した．科学分野のオントロジーとして，SWEET Ontology [*3] 内のコンセプト名から 7 個を用いた．SWEET Ontology は科学分野全体をカバーする包括的なオントロジーであるが，今回は地球科学および環境科学の分野から選択した．最近の研究動向由来のキーワードとして，Microsoft Academic Search [*4] の環境科学分野のキーワードから 6 個を選択した．

∗3 http://sweet.jpl.nasa.gov/ontology/
∗4 http://academic.research.microsoft.com/

実際の検索キーワードとして Google Trends [*5] の科学分野内のキーワードおよび Cross-DB の検索キーワード履歴から 7 個を選択した．

表 1: 評価実験用キーワード

SWEET	MSAS	GoogleTrends & Cross-DB
marine biology	global climate	atmospheric circulation
sediment	natural gas	interannual variability
acid rain	ocean current	sea level pressure
aerosol	black carbon	water quality
global warming	ozone hole	carbon cycle
air pollution	ash flow	particulate matter
southern oscillation		boreal forest

検索対象の科学データとして Pangaea が持つ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 1 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．

STT-PRF は空間/時間/テキストの全てについてクエリ拡張を行えるが，一部についてのみ行うことも可能である．評価実験では時間情報のみクエリ拡張を行う場合 (T-PRF)，空間情報のみの場合 (S-TRF)，時空間情報の場合 (ST-PRF) の 3 種類について性能評価を行い，クエリ拡張を行わない場合 (baseline) の結果と比較して時空間情報を活用することの有効性を示す．クエリ拡張に用いる仮適合データセット数 L は 10 とし，空間情報および時間情報に対する重み w_s，w_t は予備実験の結果からそれぞれ 0.370 および 0.074 とした．さらに実験に用いるキーワード毎のデータセット集合に対し，一定割合でデータセットの概要情報の削減を行う．本稿では削減率を 100%，99%，98%，95%，90%，80%，50% および 0% とした．例えば削減率 100% はデータセット集合中の全データセットから概要が削除されていることを意味し，この場合残るテキスト情報は題目及び著者名のみとなる．

本稿では性能評価に用いる指標として nDCG@30，Precision@30(P@30)，Recall@30(R@30) 及び Average Precision (AP) を用いた．nDCG はランク付けされた検索結果が関連度順にどれだけ並んでいるかを表し，上位 n 件から求める nDCG@n は式 (8) によって定義される．

$$\begin{aligned} \mathrm{nDCG@}n &= \frac{\mathrm{DCG@}n}{\mathrm{IDCG@}n}, \\ \mathrm{DCG@}n &= r_1 + \sum_{i=2}^{n} \frac{r_i}{\log_2 i}. \end{aligned} \tag{8}$$

ここで r_i はランキング第 i 番目のデータセットの関連度を表し，IDCG@n はランキングが関連度の大きい順に並んだ理想的な場合の DCG@n を表す．同様に検索結果上位 n 件での P@n および R@n はそれぞれ以下の式で与えられる．

$$\mathrm{P@}n = \frac{tp@n}{tp@n + fp@n}, \tag{9}$$

$$\mathrm{R@}n = \frac{tp@n}{tp@n + fn@\mathrm{ALL}}, \tag{10}$$

ここで $tp@n$ および $fp@n$ はそれぞれ上位 n 件までの真陽性，偽陽性を表す．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 は全件での偽陰性を表す．Average Precision は式 (11) で定義される．

$$\mathrm{AP} = \frac{1}{N} \sum_{n=1}^{N} \mathrm{rel}(n)\mathrm{P@}n, \tag{11}$$

ここで rel(n) は n 番目のデータセットがキーワードに関連すれば 1 を，しなければ 0 を返す関数である．

3.2 性能評価

表 2 に概要削減率を変化させた場合の 20 キーワードに対する性能評価指標の平均値を示す．表中の ave. #hit は平均データセット検出数である．ST-PRF は概要が 50% 以上削減されている場合，平均データセット検出数，R@30，AP において baseline を上回る性能を示す．これはテキスト情報が不足しがちな科学データ検索において，時空間情報を用いたクエリ拡張を行うことでテキスト情報では対象外とされたデータを発見することが可能となったことを意味する．一方で nDCG@30 および P@30 は baseline に及ばない場合もあるが，その差は大きくなく，また ST-PRF がクエリ拡張のための手法でありリランキングは目的としていないためでもある．

表 2: 各手法の性能比較

手法		ave. #hit	nDCG@30	P@30	R@30	AP
100% 削減	baseline	11.64	**0.54**	**0.35**	0.07	0.08
	T-PRF	16.50	0.47	0.31	0.08	0.10
	S-PRF	16.50	0.49	0.33	0.08	**0.12**
	ST-PRF	**18.93**	0.46	0.33	**0.09**	**0.12**
99% 削減	baseline	12.29	**0.61**	0.39	0.07	0.08
	T-PRF	17.21	0.53	0.36	0.08	0.11
	S-PRF	19.93	0.59	**0.40**	0.11	**0.14**
	ST-PRF	**22.36**	0.56	0.39	**0.12**	**0.14**
98% 削減	baseline	12.71	**0.75**	0.40	0.07	0.09
	T-PRF	17.93	0.67	0.36	0.08	0.11
	S-PRF	21.07	0.63	**0.42**	0.13	0.14
	ST-PRF	**23.79**	0.60	0.41	**0.14**	**0.15**
95% 削減	baseline	14.43	**0.68**	**0.38**	0.09	0.10
	T-PRF	22.79	0.58	**0.38**	0.12	0.13
	S-PRF	28.50	0.59	0.37	0.18	0.18
	ST-PRF	**33.36**	0.58	**0.38**	**0.20**	**0.19**
90% 削減	baseline	16.79	**0.69**	0.38	0.10	0.12
	T-PRF	26.29	0.58	**0.39**	0.15	0.14
	S-PRF	32.71	0.62	0.37	0.19	0.20
	ST-PRF	**37.57**	0.61	0.37	**0.20**	**0.21**
80% 削減	baseline	22.50	**0.73**	0.45	0.16	0.16
	T-PRF	34.14	0.67	0.45	0.19	0.18
	S-PRF	37.43	0.71	**0.46**	0.22	0.23
	ST-PRF	**45.00**	0.70	**0.46**	**0.23**	**0.24**
50% 削減	baseline	38.79	0.82	**0.55**	0.35	0.29
	T-PRF	51.36	0.80	0.47	0.33	0.33
	S-PRF	51.00	**0.84**	0.53	0.40	**0.38**
	ST-PRF	**59.86**	0.82	0.50	**0.41**	**0.38**
0% 削減	baseline	66.43	**0.82**	**0.60**	**0.51**	0.53
	T-PRF	76.43	0.79	0.53	0.49	0.54
	S-PRF	71.36	**0.82**	0.53	0.50	**0.56**
	ST-PRF	**79.00**	0.81	0.50	0.50	0.55

また，S-PRF の性能は T-PRF を上回る傾向にある．これは検索対象であるデータセット集合の時空間情報保有率に依存する

*5 http://www.google.com/trends/

と考えられる．例えばキーワードが “atmospheric circulation” の場合，全 120 データセット中時間情報を持つものは 33 件であるのに対し，空間情報をもつものは 115 件と大きく異なる．これは Pangaea から得られたデータセットに共通する傾向であり，その他のキーワードに関しても同様であった．これらのことから PRF の性能が検索対象のテキスト情報保有率に依存することと同じように，ST-PRF は検索対象の時空間情報保有率に依存することがわかる．

3.3 空間クエリ拡張例

図 2 から 4 に空間クエリ拡張による検索結果の変化例を示す．図 2 はキーワード “natural gas “ に対するデータセット集合の空間分布を表す．図中の緑丸は関連のあるデータセットの中心点を，赤菱は関連のないデータセットの中心点を表す．このとき，前者は主に北アメリカに分布し，後者はそれ以外の場所に分布する傾向がみられる．図 3 はテキスト情報による 1 回目の検索結果で得られたデータセットの分布を示す．検索結果として関連しないデータセット 1 件を含む 5 件が得られ，これを空間クエリとして 2 回目の検索が行われる．1 回目の検索結果でキーワードに関連する結果が空間クエリに多く含まれており，これらに対して空間スコア ϕ_s が高いものが 2 回目の検索で得られることになる．空間クエリ追加後の検索結果を図 4 に示す．北アメリカに分布するデータセットによって，その近辺の関連するデータセットが新たに検索されていることが分かる．

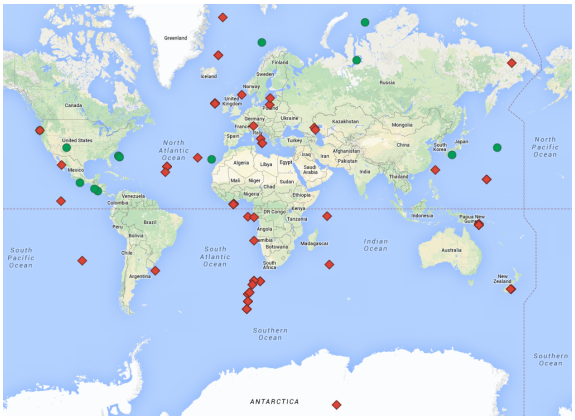

図 2: ”natural gas” の検索対象データセット集合

4. まとめ

本稿では検索対象のテキスト情報を用いてクエリ再構築，再検索を行う疑似適合性フィードバックに対し，クエリ再構築の対象を時空間情報へと拡大した STT-PRF を提案した．また，そのために必要となる時空間情報に基づくデータセット間の距離を，正規分布間のバタチャリヤ距離に基づいて定義した．評価実験によって，テキスト情報が不十分な場合における提案手法による時空間情報を用いた再検索の有効性が示された．

今後の課題としては，時空間距離定義を活用したデータセットのクラスタリング，時空間以外の情報を用いた疑似適合性フィードバック (X-PRF)，Cross-DB への応用による異分野検索時の傾向の調査などがある．

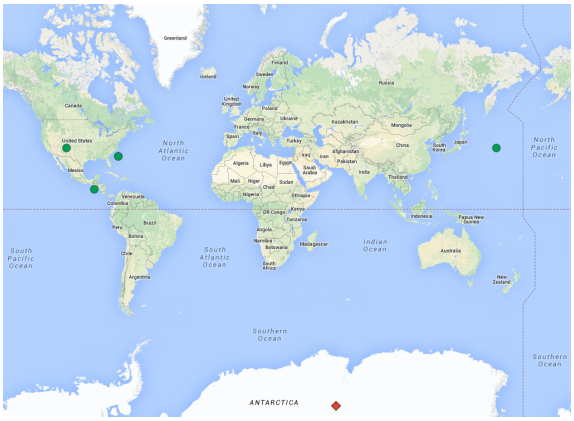

図 3: テキスト情報のみによる検索結果 (=空間クエリ)

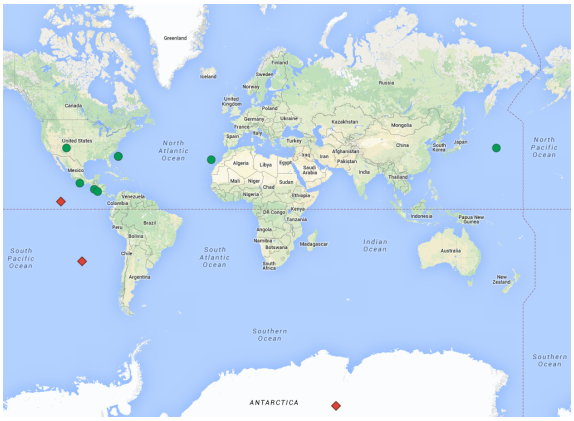

図 4: 空間クエリ追加後の検索結果

参考文献

[1] Tony Hey, Stewart Tansley, and Kristin Tolle (eds.), “The Fourth Paradigm: Data-Intensive Scientific Discovery”, Microsoft Research, 2009.

[2] I. Ruthven and M. Lalmas, “A survey on the use of relevance feedback for information access systems,” *The Knowledge Engineering Review*, Vol. 18, No. 2, pp. 95–145. Jun. 2003.

[3] C. Carpineto and G. Romano, “A survey of automatic query expansion in information retrieval,” *ACM Computing Surveys*, Vol. 44, No. 1, pp.1:1–1:50, Jan. 2012.

[4] C. Buckley, G. Salton, and J. Allan, “Automatic retrieval with locality information using SMART,” in *Proc of the 1st Text Retrieval Conference (TREC-1)*, pp.59–72. 1992.

[5] C. Lioma, M. F. Moens, and L. Azzopardi, “Collaborative annotation for pseudo relevance feedback,” in *Proc. of the ECIR’08 Workshop onf Exploiting Semantic Annotations in Information Retrieval*, pp.25–35. 2008.

[6] L. Chen, L. Chun, L. Ziyu, and Z. Quan, “Hybrid pseudo-relevance feedback for microblog retrieval,” *Journal of Information Service*, Vol. 39, No. 6, pp.773–788. Dec. 2013.

[7] S. Whiting, I. A. Klampanos, and J. M. Jose, “Temporal pseudo-relevance feedback in microblog retrieval,” in *Advances in Information Retrieval*, ser. Lecture Notes in Computer Science, Vol. 7224, pp.522-526. 2012.

[8] Eloy Gonzales, Bun Theang Ong, and Koji Zettsu, “Searching Inter-disciplinary Scientific Big Data based on Latent Correlation Analysis,” in *Proc. of the IEEE International Conference on Big data*, pp. 6–9. Santa Clara, US. Oct. 2013.

[9] A. Bhattacharrya, “On a measure of divergence between two statistical populations defined by their probability distributions”, Bulletin of the Calcutta Mathematical Society Vol. 35, pp.99–109. 1943.