IEEJ Transactions on xx
Vol.131 No.1 pp.1–8 DOI: 10.1541/ieejxxs.131.1

Paper

Spatio-Temporal Pseudo Relevance Feedback for Scientific Data Retrieval

Shin'ichi Takeuchi*a) Non-member, Komei Sugiura* Non-member
Yuhei Akahoshi* Non-member, Koji Zettsu* Non-member

We consider the problem of searching scientific data from heterogeneous ocean of scientific data repositories. This problem is challenging because scientific data contain relatively few text information comparing other search targets such as web pages. On the other hand, the metadata of scientific data contain other characteristic information such as spatio-temporal information. Although using these information has possibility to improve the search performance, many widely adopted scientific data search engines use these information only for narrowing down the search results. In this paper, we propose a novel query generation method using spatial, temporal, and text information based on pseudo relevance feedback. The proposed method generates new spatio-temporal queries from initial search results. By using these queries, the search results are re-ranked so that more related results can obtain higher ranks. The experimental results show that the proposed method outperforms a baseline method when search targets do not have rich text information.

Keywords: pseudo relevance feedback, spatio-temporal and text information, information retrieval, scientific data, query generation

1. Introduction

Data-driven science, or e-Science, is a new paradigm that goes further than mere experimental and theoretical research, and computer simulation [1]. In this paradigm, scientific data, consisted by observations and results of scientific activities, are shared and re-used so that scientists can accelerate their research activities. Free and open access to publications and scientific data provided by publicly funded research offers significant social benefits.

This has led to an explosion in the availability of scientific datasets [2], including the raw data directly extracted by measuring instruments and also the derived data from computational models and simulations [3]. These datasets can be stored on-line in large volume in public or private repositories and made accessible to users within a scientific community or beyond to foster interorganizational and inter-disciplinary research that can accelerate scientific discovery [4], [5]. Such published datasets, which number in the millions, continue to grow at an impressive rate and are long-term archived in affordable cloud storage and on disks [6].

For a scientist, discovering an appropriate set of datasets is critical for computations and effective experimental simulations. However, up to now, the selection process has typically been haphazard: researchers recycle datasets with which they are familiar, or that they have heard about through word-of-mouth. No good tools are available for dataset discovery because no widely used resources indexes them.

Several search engines such as World Data System (WDS) [†] and Pangaea [††] have been designed for discovering scientific datasets. Although these search engines are widely used, millions of datasets are searched by an engine that retrieves and ranks them using a simple keyword-based matching algorithm and by aggregating the results. One drawback with such approaches is that the accuracy of the results is largely dependent on the ability of users to formulate queries by keywords.

In this paper, we describe a novel query generation method for searching scientific datasets that performs the following two crucial functions:

- extends conventional text-based Pseudo Relevance Feedback methods by using space and time information (Section 4.1),
- scores the space and time distances by the Bhattacharyya distance among datasets and uses this information to rank the search results (Section 4.2).

2. Related Work

2.1 Scientific Data Repository At first, we clarify the definition of a dataset. A dataset is consisted by raw data and metadata. Raw data are consisted by sets of observations or results of scientific activities. Metadata represent several features of raw data(e.g. spatial, temporal, and other text information such as author name, citation, abstract, observed parameter, etc.) Figure 1 shows an example of dataset description. This dataset has spatial, temporal, and text information such as title, author, abstract, etc. This information is used as metadata of the dataset. The variety of information in metadata of a dataset is depend on the domain of dataset and the data repository.

a) Correspondence to: s.takeuchi@nict.go.jp
* National Institute of Information and Communications Technology
3-5 Hikaridai, Seika-cho, Soraku-gun, Kyoto, 619-0289, Japan

† http://www.icsu-wds.org/
†† http://www.pangaea.de/

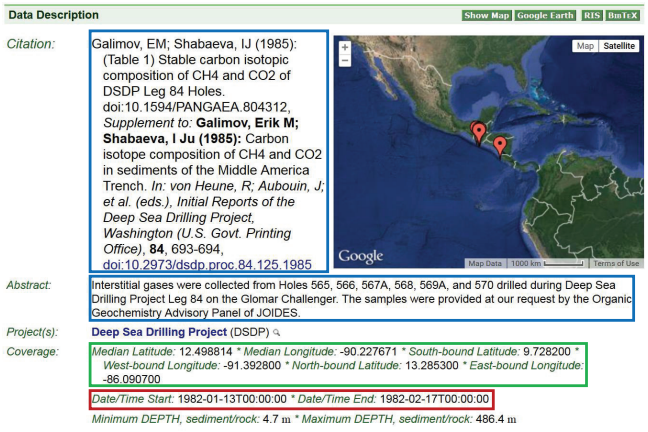


Figure 1. An example of dataset description. This description contains spatial, temporal, and several piece of text information such as title, author, abstract, etc. This information is used as a metadata of the dataset. The variety of information in the metadata is dependent on the domain of the dataset and the data repository.

Finding datasets for scientific work requires discovering the right piece of data across a wide range of interrelated scientific domains, including interdisciplinary research about a particular natural phenomenon. By using the current systems or portals for searching for datasets, if a user lacks prior knowledge about the data attributes and terminology, discovering relationships among them is very difficult.

Several search engines have been designed for discovering scientific datasets. One of the biggest scientific data repositories is the World Data System (WDS), which provides a portal for searching through a huge amount of scientific datasets [†]. The search results are ranked by its relevance to the input query.

WDS incorporates scientific data from more than 100 stand-alone data centers. Quandl[††], another search engine for numerical data, focuses on financial, economic, and social datasets and currently indexes more than eight million pieces of data.

Several repositories are consisted by datasets with rich text information and another has different tendency. Table 1 shows information existence ratio of Pangaea. Only about 1.7% of datasets have abstract but 73.2% of datasets have both space and time information. Here, we define the existence ratio of each information type. For example, the existence ratio for abstract, space, and time are denoted as R_a, R_s, and R_t and calculated as shown in Table 1.

2.2 Pseudo Relevance Feedback Dataset search systems of conventional scientific dataset portal are mainly based on text-based retrieval using metadata of datasets. However, exact match of text information is not enough for dataset search. For example, when we search "global warming" in WDS which is one of the data portal for earth science, the number of result is only 130. There must be more datasets conceptually related to global warming, but they do not exactly match "global warming". Such kind of problems are caused by the difference between index (made from metadata of datasets) and query (given by user intension.)

[†] http://www.icsu-wds.org/services/data-portal
[††] http://www.quandl.com/

Table 1. Information existence ratio in the metadata of scientific datasets in Pangaea. All datasets were harvested on January 23, 2014.

datasets	num. of datasets	ratio
overall	405,456	
w/ abstract	7,028	$R_a = 0.017$
w/ time info.	297,478	$R_t = 0.733$
w/ space info.	404,145	$R_s = 0.996$
w/ space and time info.	297,037	$R_{st} = 0.732$

To find more related datasets, relevance feedback [(7)] is used in information retrieval. The notion of relevance feedback is an iterative process, where users first specify which documents are relevant for them. These specified documents are used by the system to retrieve more or similar documents. From the newly retrieved documents, the users once again specify relevant documents to produce a new query. The process is then repeated. When the user's interactions are removed from this iterative process, this kind of relevance feedback is called Pseudo Relevance Feedback (PRF) or blind relevance feedback [(8)], [(9)]. The idea is to perform a normal search and assume than the top k ranked documents are relevant. Then, using this information, query expansion is performed to retrieve more similar candidate documents.

Different variations of PRF have been explored and applied to specific problems. In Lioma's work [(10)], queries are expanded using the semantic annotations found in collaborative tagging systems. In the context of microblog retrieval, Chen [(11)] and Whiting [(12)] introduced a dynamic PRF that extracts representative terms based on the query's temporal profile. They show that exploiting temporal evidence for microblogs is effective. In Yin's work [(13)], spatial relationships are used instead of iconic image retrieval. A specific data structure that allows region-based feedback improves the system's efficacy.

3. The Task: Searching Scientific Dataset

Although aforementioned search engines are widely used, millions of datasets are searched through an engine that retrieves and ranks them using a simple keyword-based matching algorithm and by aggregating the results. One drawback with such approaches is that the accuracy of the results is largely dependent on the ability of the users to formulate queries by appropriate keywords. Several frameworks to manage scientific datasets with spatio-temporal information have been proposed [(14)], [(15)].

There are several well-known techniques to improve the accuracy of the search results such as interactive query refinement [(16)], information filtering [(17)], word sense disambiguation [(18)], search results clustering [(19)]. In the particular context of scientific dataset discovery, most currently available search engines use some variation of query expansion/text mining, clustering, or semantics [(8)], [(20)], [(21)]. Another trend in dataset search crawls the hidden web and efficiently indexes metadata that provide the basic building blocks for sifting through this vast space [(22)], [(23)], [(24)].

In this paper, we apply PRF to large-scale open scientific data search system. For the scientific data retrieval, to use the

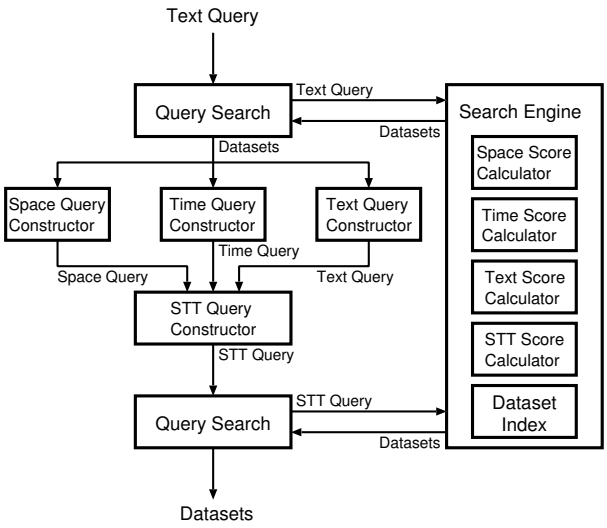


Figure 2. The architecture of STT-PRF. Conventional PRF only uses a Text Query Constructor for additional text query. In contrast, STT-PRF additionally possesses a Time Query Constructor and a Space Query Constructor for additional time and space query.

metadata of the dataset is more important. Text information, one of the metadata of datasets, is important information to search datasets. Also, when a phenomenon is observed at some time and some place, the phenomena observed nearby the place and time are considered as related. We apply this intuitive approach to conventional PRF. To find the datasets related to such concept, not only their text information but also space/time information affects the relevance of dataset. Actually, the importance of spatio-temporal data mining is growing in many domains such as climatology, marine ecosystem, traffic control and so on [25].

4. Proposed Method

This section describes in detail our proposed method, named Spatio-Temporal and Text Pseudo Relevance Feedback (STT-PRF). STT-PRF is based on PRF but its particularity is that it not only uses text information but also spatio-temporal information.

4.1 PRF using Spatio-Temporal Information Figure 2 shows STT-PRF's architecture. First, with the system input layer, the user inputs a search keyword in the form of a text string. With the Query Search component, a standard text-based search algorithm is applied using the text information included in each dataset's metadata. The retrieved datasets are then ranked by their text scores ϕ_k calculated by the Text Score Calculator. In this paper, ϕ_k is given by the cosine distance between the TF-IDF based feature parameter from the keyword and a dataset's text information. Using the cosine distance is one of the most standard techniques to represent the similarity between two documents. However, we stress that STT-PRF's comparative results are independent from those choices. Since we are working toward a more domain-optimized method, we will eventually replace the current distance measures by more adequate techniques.

Query generation is the subsequent step. In a standard PRF, the algorithm treats the top L datasets in the initial ranking, referred to as Y_L, as relevant datasets. Then the Text Query Constructer component makes additional text queries from the text information. In STT-PRF, however, the query is not solely composed of text information but also of space and time information. The beginning and end dates of each dataset in Y_L form time queries by the Time Query Constructor component. Similarly, their spatial coverage is comprised of space queries that are built by the Space Query Constructor component. The set of the text, time and space queries is treated as an expanded query and is used by the second Query Search component.

With the second Query Search component, the space and time scores are calculated for each dataset by the Space Score Calculator and Time Score Calculator. They compute space score ϕ_s and time score ϕ_t for dataset y using the following equations:

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

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

Here, y' shows the dataset in Y_L, which is the set of datasets treated as relevant ones. d_s and d_t stand for the space and time distances between the two datasets, respectively, as described in Section 4.2.

The space, time and text scores of all the indexed datasets are calculated, and the total score of dataset y, written $\phi(y)$, is given by:

$$\phi(y) = w_s\phi_s(y) + w_t\phi_t(y) + \phi_k(y). \quad (3)$$

where w_s and w_t are the weight parameters for each distance. After the score calculation, the indexed datasets are ranked based on their total scores and the second Query Search step outputs the ranked datasets.

Note that we use keywords (text information) as a input query which is the standard interface of search system. STT-PRF can be applied with other (spatial or temporal) input query.

4.2 Distance between Datasets Although PRF uses text information for the second Query Search step, STT-PRF additionally uses space and time information and calculates the distance between two datasets to obtain the space and time scores. There are several definitions of spatial/temporal distance (or similarity) [26]. As a previous research, the authors investigated the effectiveness of the STT-PRF with different definition of datasets and limited test sets [27]. In this paper, these distances are used in Eqs.(1) and (2). In this section, we describe these distances in detail using the information based on a metadata of datasets.

4.2.1 Spatial/Temporal Information in Metadata A dataset's space and time information are present in its metadata and are defined as one or two dimensional range. Such information is specified under the form of beginning point $\boldsymbol{x}_b$ and end point $\boldsymbol{x}_e$ within a time or a spatial series.

For the temporal information, the beginning and end times become the beginning and end points of a time series. For example, the beginning and end points of a dataset that starts at year 1990 and ends at year 2000 has $(\boldsymbol{x}_b, \boldsymbol{x}_e)$ set as $(1990, 2000)$.

For the space information, the north/south points correspond to the beginning/end points in the latitude series, and the west/east points indicate the beginning/end points of the longitude series.

4.2.2 Distance Definition based on Bhattacharyya Distance Here, we explain the definition of space and time distances among datasets based on the Bhattacharyya distance [28]. The Bhattacharyya distance is one of the standard definition to measure the distance between two probability distributions and defined as follows:

$$d_B(p,q) = -\ln\left(\int \sqrt{p(x)q(x)}dx\right). \qquad (4)$$

Here, p and q are some probability distribution. But as mentioned in previous section, the spatial/temporal information is given as one or two dimensional range. Therefore, we approximate these information using the normal distribution to apply Bhattacharyya distance for the distance of datasets.

For the approximation of datasets by the normal distribution, the uniform distribution is considered as the first step. The uniform distribution is simply given by beginning point $\boldsymbol{x}_b$ and end point $\boldsymbol{x}_e$ of spatial/temporal information. The advantage of this consideration is that the mean $\boldsymbol{\mu}$ and variance Σ of the distribution is defined as follows:

$$\boldsymbol{\mu} = \frac{1}{2}(\boldsymbol{x}_e + \boldsymbol{x}_b), \quad \Sigma = \frac{1}{12}(\boldsymbol{x}_e - \boldsymbol{x}_b)^2. \qquad (5)$$

After that, we can consider a normal distribution which has same mean and variance given by Eq.(5) as an approximation of spatial/temporal information of a dataset. The distance between datasets is defined as the Bhattacharyya distance between normal distributions whose approximate target datasets. Especially, the Bhattacharyya distance for normal distributions are further transformed as follows:

$$d(y_i, y_j) = \frac{1}{8}(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)^\top \left[\frac{1}{2}(\Sigma_i + \Sigma_j)\right]^{-1} (\boldsymbol{\mu}_i - \boldsymbol{\mu}_j) + \frac{1}{2}\ln\left\{\frac{\det(\frac{1}{2}(\Sigma_i + \Sigma_j))}{\sqrt{\det(\Sigma_i)\det(\Sigma_j)}}\right\}. \qquad (6)$$

5. Experiments

5.1 Experimental Setup The performance of keyword based search is largely dependent on the ability of users to formulated good queries. Therefore the performance should be evaluated with commonly used queries. At this point, we use keywords from actual query lists which is obtained from the famous search engines. For the performance evaluation of STT-PRF, 50 scientific keywords are chosen from major search engines. The keywords were chosen from actual science related keywords obtained from Google Trends [†] and the query logs of the Cross-DB search system. Additional keywords were chosen from environmental science fields using Microsoft Academic Search [††] for the current trends in searched for terms. More keywords were chosen from ontological concepts and were created using the SWEET ontology, which mainly covers the earth and environmental science terms. These keywords were selected from natural science domains. Table 2 presents examples of the keywords.

Table 2. Keywords for evaluation experiments chosen from queries of major search engines and additional sources.

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	

For these experiments, we took datasets from Pangaea by searching with keywords shown in Table 1. All of keywords brought more than 120 dataset as a search results so that we used top 120 datasets of them. Keyword based search results may contain noise datasets (not so keyword-related datasets.) Therefore we used top 100 and additional 20 datasets to get more keyword-related datasets in lower rank. The relevance of all the retrieved datasets was manually evaluated by three human labelers with master's degrees in natural science. The relevance of the retrieved datasets according to the queries were evaluated on a scale from 0 to 3. A dataset with a relevance value of 3 is completely related to the target query. A relevance value of 0 means that it is completely unrelated to the query. In the following experiments, datasets with relevance values of 2 or 3 were considered *query-related* [†††].

Weight parameters w_s and w_t in Eq.(3) were set to 0.370 and 0.074. These weights are set to optimal values based on preliminary experiments. The top ten datasets in the initial ranking were also used as relevant datasets ($L = 10$). For our experiments, since the amount of available data was limited, we empirically determined the values of weight parameters w_i in Eq.(3).

It is important to investigate the performance of STT-PRF under various conditions. One of the important condition of search target repository is R_a because it influences the 1st search step of STT-PRF. Therefore, we evaluate the performance of STT-PRF with test datasets changing R_a. The aim is to evaluate the robustness of the methods against the lack of the text information. As shown in Table 1, the R_a of original Pangaea is about 0.017. In the experiments, the R_a had the following values: 0.01, 0.02, 0.05, 0.10, 0.20, 0.50, and 1.00. Note that R_a is small for actual scientific data. For example, $R_a = 0.017$ for datasets in Pangaea as described in Table 1. In other words, the results from large R_a are not so important. Considering the actual use case, we examined mainly small R_a. The result with $R_a = 1.0$ is examined to grasp the whole

[†] http://www.google.com/trends/

[††] http://academic.research.microsoft.com/

[†††] http://www2.nict.go.jp/univ-com/isp/s.takeuchi/sttprf.tgz

Table 3. Comparison of performance of various methods with different R_a. The combination of Normal distribution and Bhattacharyya distance outperforms other combinations in most of conditions.

R_a	Distribution	Distance	ave. #hit	P@30	R@30
0.01	Normal	Bhattacharyya	87.7	**0.328**	**0.219**
	Normal	L_2	**88.4**	0.321	0.206
	Uniform	Bhattacharyya	88.3	0.315	0.201
	Uniform	L_2	88.1	0.324	0.207
0.05	Normal	Bhattacharyya	95.7	**0.332**	**0.261**
	Normal	L_2	**96.3**	0.324	0.248
	Uniform	Bhattacharyya	**96.3**	0.322	0.246
	Uniform	L_2	96.1	0.322	0.236
0.20	Normal	Bhattacharyya	102.5	**0.356**	**0.328**
	Normal	L_2	**106.6**	0.346	0.306
	Uniform	Bhattacharyya	**106.6**	0.343	0.293
	Uniform	L_2	104.8	0.344	0.293

tendency of the effectiveness of the proposed method.

Although the main effectiveness of PRF is to improve recall performance, we use the following performance evaluation criteria: Precision at 30 (P@30) and Recall at 30 (R@30). The Precision of the top n datasets P@n and the recall of the top n datasets R@n are defined as:

$$\mathrm{P@}n = \frac{tp@n}{tp@n + fp@n}, \quad \mathrm{R@}n = \frac{tp@n}{tp@n + fn@\mathrm{ALL}}, \tag{7}$$

respectively, where $tp@n$ is the number of true positives and $fp@n$ is the number of false positives in the top n datasets. In our experiments, since the number of query-related datasets is known as N, the false negatives can be calculated. We denote $fn@$ALL as the false negatives from all of the datasets.

In general, web page search methods are mainly evaluated by using top 10 results of them because the user often check and obtain the information from only the first pages. They do not need so many matched pages because they have only to find the first appropriate one. On the other hand, searching scientific datasets has different tendency. Although the most relevant dataset is required, other related datasets are also useful to compare or support the first one. At this point, we evaluate top 30 results based on the assumption of that each result page shows 10 results and user checks three result pages.

As described in Section 4, we use the combination of Bhattacharyya distance and normal distribution for the definition of distance between datasets. To show the validity of proposed method, we also use other definitions.

For the comparison of distance measure, we use L_2 distance given by Eq.(8)

$$d_{L_2}(p, q) = \int (p(x) - q(x))^2 dx. \tag{8}$$

Also, the uniform distribution is used as the comparison of distribution definition.

5.2 Performance Evaluation Table 3 shows the average number of hit datasets (ave. #hit) and the performance measures for different R_a and the combination of the distribution and distance to represent spatio-temporal information of a dataset. This result reveals that the combination of the normal distribution and Bhattacharyya distance outperforms other combinations almost outperforms in P@30 and R@30.

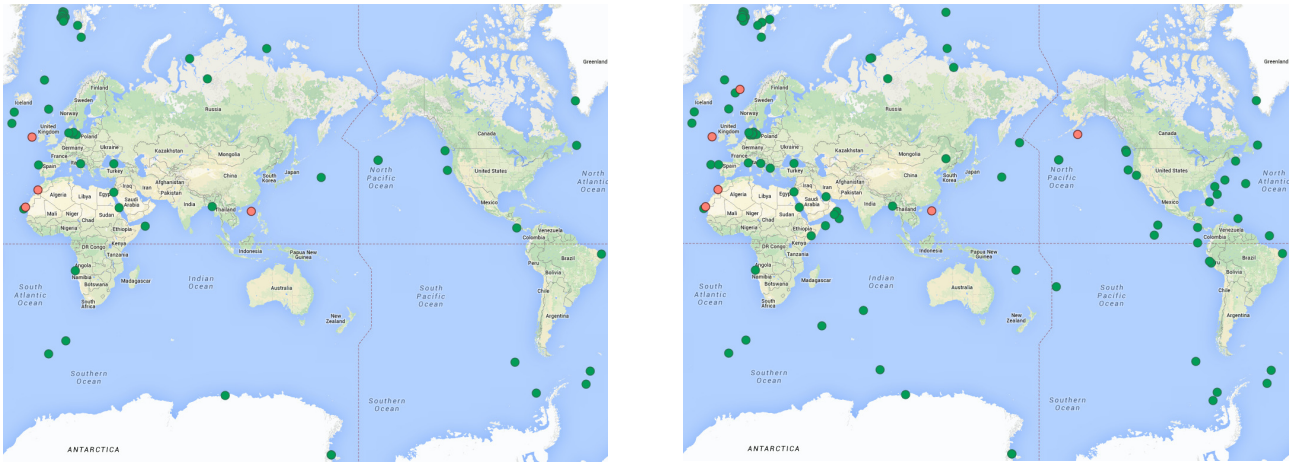
Table 4. Performance comparison among various methods using different R_a. The value of R_a of Pangaea is about 0.017.

R_a	Method	ave. #hit	P@30	R@30
0.01	baseline	14.6	**0.366**	0.090
	S-PRF	19.7	0.332	0.109
	T-PRF	20.6	0.358	0.115
	PRF	87.4	0.329	0.202
	STT-PRF	**87.7**	0.328	**0.219**
0.02	baseline	15.0	**0.388**	0.095
	S-PRF	21.4	0.357	0.126
	T-PRF	21.3	0.370	0.123
	PRF	91.5	0.332	0.221
	STT-PRF	**91.6**	0.332	**0.238**
0.05	baseline	16.7	**0.395**	0.115
	S-PRF	26.3	0.360	0.180
	T-PRF	24.5	0.372	0.152
	PRF	95.5	0.331	0.240
	STT-PRF	**95.7**	0.332	**0.261**
0.10	baseline	19.1	**0.434**	0.136
	S-PRF	30.4	0.399	0.196
	T-PRF	27.7	0.422	0.179
	PRF	101.4	0.349	0.260
	STT-PRF	**101.8**	0.336	**0.278**
0.20	baseline	24.0	**0.464**	0.208
	S-PRF	36.1	0.427	0.259
	T-PRF	33.1	0.451	0.244
	PRF	102.0	0.359	0.311
	STT-PRF	**102.5**	0.356	**0.328**
0.50	baseline	38.6	**0.497**	0.336
	S-PRF	49.2	0.462	0.354
	T-PRF	49.0	0.465	0.345
	PRF	107.7	0.391	0.388
	STT-PRF	**108.0**	0.395	**0.398**
1.00	baseline	62.9	**0.525**	**0.433**
	S-PRF	68.7	0.467	0.407
	T-PRF	71.4	0.469	0.428
	PRF	**113.3**	0.402	0.417
	STT-PRF	**113.3**	0.398	0.417

Table 4 shows the same criteria for STT-PRF given by the combination of the normal distribution and Bhattacharyya distance. In this experiments we used a keyword-based search without any PRF as a baseline method. However it is possible to apply other search result improvement methods described in Section 3, this paper focuses on the efficiency of using spatio-temporal information for PRF. At this point, we used PRF using spatial information, PRF using temporal information, and conventional PRF (PRF using text information) as comparison methods. In Table 4, those methods are denoted by S-PRF, T-PRF, and PRF respectively.

The results reveal that STT-PRF outperforms R@30 and the number of hit datasets with most of abstract existence ratio. STT-PRF can bring in additional datasets that cannot be found by text-based search only methods. Although not all of the PRF methods show improvement of Precision, their main focus is not quality but the amount of search results.

More specifically, due to the experiment design, the PRF result is close to the STT-PRF result. All of the test sets were corrected based on the text search result, and the datasets are highly correlated not by spatial/temporal information but by text information. However, STT-PRF outperforms other methods, and this result shows the effectiveness of using spatial/temporal information as additional queries.



(a) Initial search results **(b)** search results with STT query generated by STT-PRF

Figure 3. **(a)** Locations of datasets found by initial query search with keyword "sediment". Green and pink circles represent the center of the relevant and irrelevant datasets respectively. **(b)** Locations of datasets found by STT query generated by STT-PRF. We can obtain additional relevant datasets.

6. Discussion

6.1 Example Of STT Query As an example of the effectiveness of spatial query, Figure 3 shows the spatial distributions of the datasets given by a search query for "sediment" with $R_a = 0.02$. In both figures, datasets with relevance of 0 or 1 (irrelevant) and 2 or 3 (relevant) are plotted with pink and green circles respectively. In the first step of STT-PRF, we did a text-based search and found 63 datasets. Figure 3 **(a)** shows the locations of the 59 relevant and the four irrelevant datasets. STT-PRF performs spatio-temporal queries starting from those locations and searches for additional datasets near those datasets. For example, the datasets near the spatial queries take higher spatial score ϕ_s. Figure 3 **(b)** shows the search result after STT-PRF. The number of relevant and irrelevant datasets increases to 109 and 6 respectively.

6.2 Cross-DB Search System Finding relationship from datasets from different domain, such as co-location pattern mining [30] has been investigated. We previously applied our proposed method to the Cross-DB Search System [29] whose specificities are: 1) it is especially oriented to discover datasets, facilitating data access and usability, 2) it provides high quality-ensured data and information, especially for scientists since it can find related and correlated datasets by integrating them using spatio-temporal relationships and information from citations and ontologies. The spatio-temporal associations are calculated from the observed data, and the ontological associations are based on the relationships between concepts on a given ontology where the datasets are represented; the citational associations leverage the use and the re-use of datasets by acting as a bridge between scientific domains.

Figure 4 shows **(a)** initial search results and **(b)** the results after STT-PRF of Cross-DB Search System for the keyword "precipitation". In Figure 4 **(a)**, datasets which actually have the keyword "precipitation" are found. Several datasets in the top of this rank is considered as a query-related datasets and used for STT-PRF. In Figure 4 **(b)**, additional datasets are found by using the STT information of query-related datasets. The number of datasets increases more than 40 times by using STT-PRF.

6.3 Optimizations Equation (3) is a linear combination, which is a standard method that combines several scores to determine the total relevance/score [31], [32]. However, in our paper we intentionally adopted a linear function for combining the scores. Although Eq.(3) currently only consists of weight parameters, we are planning to model it as an adaptive model that will preliminary training by using the Minimum Classification Error (MCE) [33] technique to determine the optimal parameters that minimize the empirical error rate. To approximate the scale of the maximum margin, we believe that a linear combination is acceptable, as shown in Eq.(3).

An alternative measurement of the distance among datasets such as KL divergence, Hellinger distance, etc. can be applied chosen by considering the target domain's specifications. The spatio-temporal information of the datasets was given by their own metadata that only contain boundary information. It is possible to use more precise representation when we use all of the spatio-temporal information in the raw data.

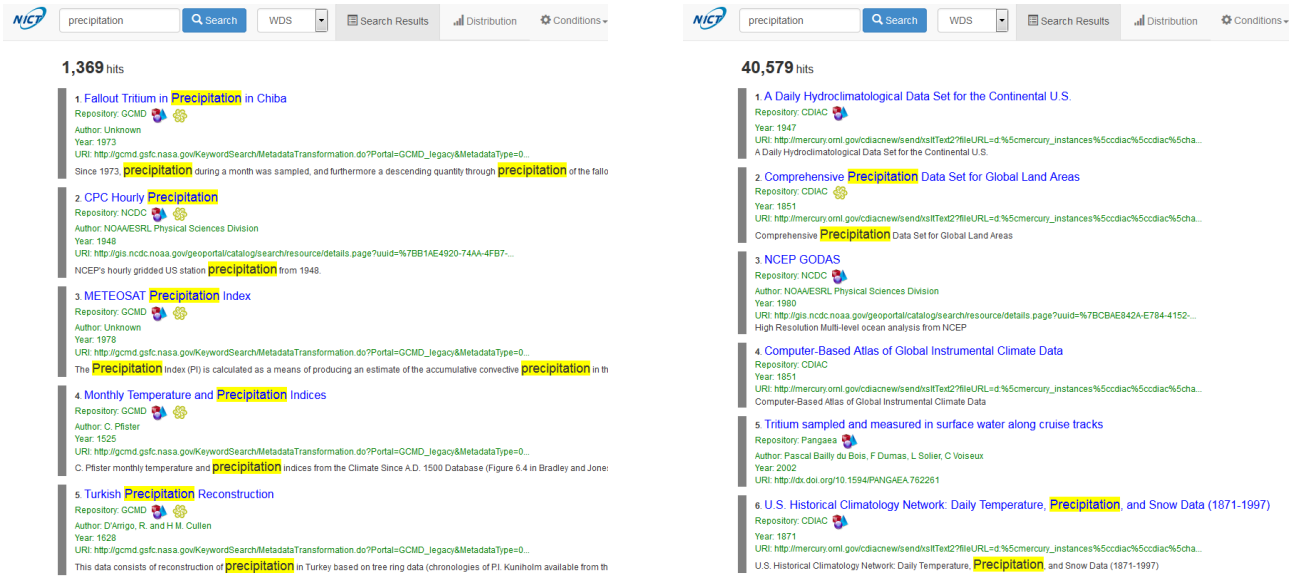
6.4 Expansion of Performance Evaluation Condition To the best of knowledge, no standard test sets exist for evaluating scientific dataset search systems, so we created our own original test set specifically for conducting the experiments described in Section 5.

Several standard evaluation test sets exist for search engines, but the difference and our motivation for creating original test sets is that the specific knowledge of a particular scientific domain is required. With that knowledge, we can assess whether a dataset is relevant to a specific query.

For our experiments, specialists created several test sets using different keywords. Therefore, the amount of test sets was constrained by financial constraints in terms of the number of specialists and time. More test sets are being created for this research work.

7. Conclusion

In this paper, we proposed a novel query generation method called STT-PRF that exploits spatio-temporal infor-



(a) Initial search results

(b) search results with STT query generated by STT-PRF

Figure 4. **(a)** Initial search results for the keyword "precipitation" by Cross-DB Search System. The datasets which actually have the keyword "precipitation" are found. Several datasets in the top of this rank is considered as a query-related datasets and used for STT-PRF. **(b)** Search results after STT-PRF. Additional datasets are found by using the STT information of query-related datasets. The number of datasets increases more than 40 times by using STT-PRF.

mation for data search. It uses text information as well as space and time information in order to expand the initial query based on a pseudo relevance feedback approach. In STT-PRF, the distance between two datasets was defined based on their spatio-temporal information. By defining the distance many similar applications can be used (e.g., clustering, recommendation.) Our experimental results demonstrate that STT-PRF can find datasets that do not have sufficient text information for applying standard text-based PRF. We also show that the retrieved datasets are highly relevant to input queries.

Future work will test STT-PRF on other scientific domains. Pangaea specializes in environmental science datasets, and we know that other domains have different spatio-temporal needs. In addition, we only focus as metadata for spatial information but all the locations in a datasets are potentially useful. For example, computing the similarity of raw data of datasets have been investigated [34] and it can be applied as a distance measurement of datasets. At this point, we will investigate on the probabilistic representation of such location. Also, adding other types of correlations such as ontological correlation, citational correlation, etc. will be used to extend PRF. We also focus on the optimization of the weight parameters for spatial/temporal/text scores in Eq. 3 as described in Section 6.3.

References

(1) T. Hey, S. Tansley, and K. Tolle (eds.), "The Fourth Paradigm: Data-Intensive Scientific Discovery," Microsoft Research, 2009.

(2) Y. L. Simmhan, S. L. Pallickara, N. N. Vijayakumar and B. Plale, "Data Management in Dynamic Environment-driven Computational Science, in Grid-Based Problem Solving Environments," Springer, Boston, MA, USA, Jul. 2007.

(3) J. Yu and R. Buyya, "A taxonomy of scientific workflow systems for grid computing," *SIGMOD Rec.*, Vol. 34, p. 3. Sep. 2005.

(4) M. Humphrey, D. Agarwal, and C. van Ingen, "Fluxdata.org: Publication and Curation of Shared Scientific Climate and Earth Sciences Data," *In Proc. of the 5th IEEE international Conference on e-Science*. Oxford, UK. Dec. 2009.

(5) "Bits of Power: Issues in Global Access to Scientific Data," Committee on Issues in the Transborder Flow of Scientific Data, National Research Council, 1997.

(6) J. Gray, A. S. Szalay, A. R. Thakar, C.Stoughton and J. VandenBerg, "Online Scientific Data Curation, Publication, and Archiving," *In Proc. of the SPIE*, Vol. 4846, pp. 103–107. 2002.

(7) 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.

(8) 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.

(9) 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.

(10) 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.

(11) 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.

(12) 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.

(13) P. Y. Yin and C. W. Liu, "A new relevance feedback technique for iconic image retrieval based on spatial relationships," Journal of Systems and Software, Vol. 82, No. 4, pp. 685–696. 2009.

(14) S. Fiore, C. Palazzo, A. D'Anca, I. Foster, D. N. Williams and G Aloisio, "A big data analytics framework for scientific data management," in *Proc. of the International Conference on Big Data*, pp. 1–8. 2013.

(15) G. Faria, C. B. Medeiros, M. A. Nascimento, "An extensible framework for spatio-temporal database applications," in *Proc. of 10th International Conference on Scientific and Statistical Database Management*, pp. 202–205. Jul. 1998.

(16) P. Anick, "Using terminological feedback for web search refinement: a log-based study," in *Proc. of 26th annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 88–95, 2009.

(17) U. Hanani, B. Shapira, and P. Shoval, "Information filtering: Overview of issues, research and systems," User Modeling and User-Adapted Interaction, Vol. 11, No. 3, pp. 203–259, 2001.

(18) R. Navigli, "Word sense disambiguation: A survey," ACM Computing Surveys, Vol. 41, No. 2, p. 10, 2009.

(19) H-J. Zeng, Q-C. He, Z. Chen, W-Y. Ma, and J. Ma, "Learning to Cluster Web Search Results," *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 210-217, 2004.

(20) A. Leuski and J. Allan, "Interactive information retrieval using clustering and spatial proximity," *User Modeling and User-Adapted Interaction*, Vol. 14, No. 2–3, pp. 259–288. Jun. 2004.

(21) A. Hogan, A. Harth, J. Umbrich, S. Kinsella, A. Polleres and Stefan Decker, "Searching and browsing linked data with SWSE: The semantic web search engine," *Web Semantics: Science, Services and Agents on the World Wide Web*, Vol. 9, No. 4, pp. 365–401. 2011.

(22) L. Barbosa and J. Freire, "Combining classifiers to identify online databases," in *Proc. of the 16th International Conference on World Wide Web,* ser. WWW'07 (ACM2007), pp. 431–440. NY, USA. 2007.

(23) S. L. Pallickara, S. Pallickara, M. Zupanski and S. Sullivan, "Efficient Metadata Generation to Enable Interactive Data Discovery over Large-scale Scientific Data Collections," in *Proc. of the 2nd International Conference on Cloud Computing Technology and Science,* pp. 573–580. Dec. 2010.

(24) A. Fox, C. Eichelberger, J. Hughes and S. Lyon, "Spatio-temporal Indexing in Non-relational Distributed Databases," in *Proc. of IEEE International Conference of Big Data*, pp. 291–299. CA. USA. Oct. 2013.

(25) V. Bogorny and S. Shekhar, "Spatial and Spatio-temporal Data Mining," *Proc. of the 2010 IEEE International Conference on Data Mining (ICDM),* p.1217. Dec. 2010.

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

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

(28) A. Bhattacharyya, "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.

(29) E. Gonzales, B. T. 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.

(30) Y. Huang, J. Pei and H. Xiong, "Mining co-location patterns with rare events from spatial data sets, " Geoinformatica, Vol.10, Issue. 3, pp.239–260 Sep. 2006.

(31) F. Skopik, D. Schall and S. Dustdar, "The Cycle of Trust in Mixed Service-oriented Systems," in *Proc. of the 35th Euromicro Conference onf Software Engineering and Advanced Applications*, pp. 72–79. 2009.

(32) B. Benatallah, Q.Z. Sheng and M. Dumas, "The Self-Serv environment for Web services composition,", in *Internet Computing*, Vol. 7, No. 1, pp. 40–48. Feb. 2003.

(33) B-H. Juang, W. Chou and C-H. Lee, "Minimum Classification Error Rate Methods for Speech Recognition," in *IEEE Trans. on Speech And Audio Processing*, Vol. 5, No. 3, pp. 257–265. May. 1997.

(34) J-R. Hwang, H-Y. Kang and K-J. Li, "Spatio-temporal similarity analysis between trajectories on road networks," Perspectives in Conceptual Modeling, pp.280-289. Jan. 2005.

Shin'ichi Takeuchi (Non-member) received Ph.D. degree in Computer Science and Engineering from the University of Aizu in 2008. He is a researcher of Information Services Platform Laboratory at Universal Communication Research Institute of National Institute of Information and Communications Technology (NICT), Japan. He was in Gifu University and engaged in research on speech recognition from 2008 to 2010. His current research interests include information retrieval and recommendation.

Komei Sugiura (Non-member) is a Senior Researcher at National Institute of Information and Communications Technology (NICT), Japan. He received his B.E. degree in electrical and electronic engineering, and M.S. and Ph.D. degrees in informatics from Kyoto University in 2002, 2004, and 2007, respectively. From 2006 to 2008, he was a research fellow at Japan Society for the Promotion of Science. In 2008, he joined NICT, where he is working on robot dialogue, imitation learning, machine learning, spoken dialogue system, service robots, and cyber-physical systems.

Yuhei Akahoshi (Non-member) Yuhei Akahoshi is a Technical Expert at National Institute of Information and Communications Technology, Japan. He received B.E. in informatics and mathematical science, and M.S. in informatics from Kyoto University in 2003 and 2005, respectively. His interests are databases, information retrieval, and cyber-physical computing.

Koji Zettsu (Non-member) received Ph.D. in Informatics from Kyoto University, Japan in 2005. He is a Director of Information Services Platform Laboratory at Universal Communication Research Institute of National Institute of Information and Communications Technology (NICT), Japan. He was a visiting associate professor of Kyoto University, Osaka University and Nara Institute of Science and Technology from 2008 to 2012. He was a visiting researcher of Christian-Albrechts-University, Kiel, Germany in 2009. His research interests are information retrieval, databases, data mining, and software engineering. He is a member of IPSJ, IEICE, DBSJ, and ACM.