Constrained Greedy Reranking Based on Spatial Diversity for Scientific Data Retrieval

Shin'ichi Takeuchi[1,a)] Komei Sugiura[1] Yuhei Akahoshi[1] Koji Zettsu[1]

Abstract:
We consider the challenge of spreading the diversity of search results for open scientific dataset search systems. To help discover novel scientific knowledge, search results require datasets with various information to spread their diversity. When the search system only focuses on the relevance between queries and datasets, some search results are not very useful because they are too similar. To improve the diversity of search results, we have to consider the relationship among search results and reduce the similar results. As a first step, we focus on the spatial diversity of search results and define the challenge of selecting more relevant datasets to avoid spatial overlapping. The problem of selecting non-overlapped regions from overlapped regions is challenging because it can be considered as a combinational optimization problem, which is known to be NP-hard. In this paper, we propose a novel selective reranking method for scientific data search systems. The proposed method selects and reranks search results by solving rectangular label placement problems using the spatial information of search results. The search results are reranked based on both query relevance and spatial distribution. We compare the performance of our method with another label placement problem solver based on the quality of the reranked results. The experimental evaluations show that the proposed method outperformed the other methods to which it was compared.

1. Introduction

Open data in many fields, such as economic, government, and science, are expected to fuel a new innovation. For economics, according to the trial calculations of JETRO/IPA in 2013, the business market concerned with open data will exceed 10 million dollars in Japan [1]. Publishing scientific datasets has developed into a global trend in many scientific domains. Archived data supported by public funding are gradually being published, making them available to the public [2]. To publish these data, many scientific repositories have their own dataset search systems. Along with this increasing trend of the presentation of open data, finding data highlights a critical problem; discovering surprising but highly relevant data is very challenging.

One difficult problem for searching datasets is whether their metadata contain enough information for search systems. In general, since scientific data do not have as much text information as web pages (Section 3.2), their other information such as spatial and temporal information are important for scientific data search systems. At this point of view, we have been developing an open scientific data search system named Cross-DB Search System [*1] [3].

We also have to consider the diversity of the search results because they often contain "similar" datasets that have the same spatio-temporal information. Even if these results are related to the search query, they are not desirable from the viewpoint of diversity. We focus on the spatial diversity of search results because spatial information of a dataset, which is generally given as a rectangle region, is one important feature to overview them. To select the non-overlapped regions from the overlapped regions is one way to improve the diversity of search results because their spatial regions are often unevenly distributed. Furthermore, most of these regions overlap. If all of the spatial information of the search results are shown at once, the users get rectangle regions that repeatedly overlap. We consider this situation a variant of the label placement problem. Since label placement problems are NP-hard, many heuristics algorithms have been proposed to solve them efficiently.

In this paper, we propose a novel selective reranking method considering spatial diversity for scientific data retrieval. The proposed method selects and reranks search results by solving rectangular label placement problems using the spatial information of the search results. By using the proposed method, the search results are reranked by considering both query relevance and spatial diversity. Figure 1 shows the concept of the proposed method.

The followings are the main contributions of this paper:

- We propose a novel algorithm named Spatially Constrained Greedy Selection (SCGS) that solves area-feature label placement problem. This method dis-

[1] National Institute of Information and Communications Technology (NICT), Kyoto 619–0289, Japan
[a)] s.takeuchi@nict.go.jp
[*1] Cross-DB Search System is available at http://dataeyez.org/crossdb/

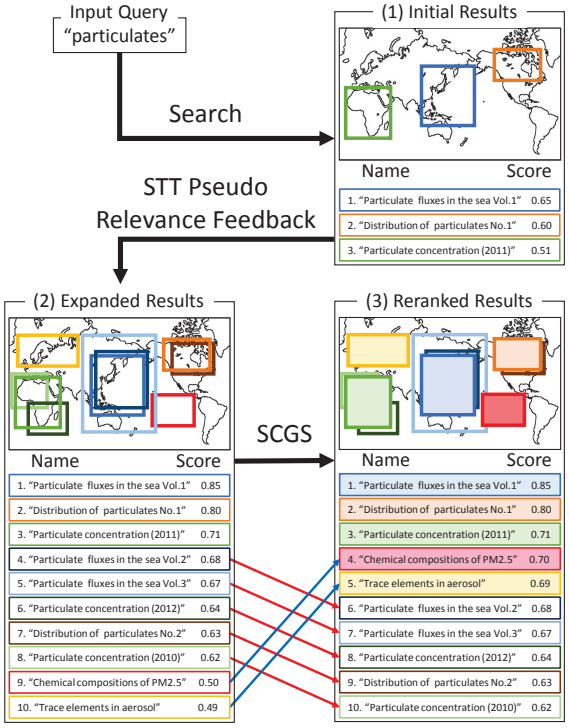


Fig. 1 The concept of search result reranking considering spatial diversity. (1) The original search results are obtained by using input query. (2) Spatio-temporal query is generated from these results, and spatially/temporally related datasets are obtained by using them. But the results are consisted by spatially/temporally "similar" datasets. (3) Search results are reranked by considering spatial overlapping so that we can obtain more spatially diversified results.

tinguishes rectangular regions from other regions that overlap (Section 4).

- We apply the SCGS method to the Cross-DB Search System, which is an open scientific data search system (Section 3) to rerank the search results to increase their spatial diversity.

2. Related Work

2.1 Open Scientific Datasets

Based on its recent trend, big data is expected to create new innovation by using open data [1]. Open government data are one important source of open data. The trend of the publication of governmental data is widely spreading [4], [5]. Governments in several countries, for example in Japan, publish their data and encourage their utilization. These data are mainly provided at Data.Gov [*2] or similar repositories for individual countries. They provide open governmental data in several formats and users can make service applications with them. CKAN [*3], which is open source software to construct data catalog sites, provides several ways to harvest raw data and the metadata of dataset. Many open governmental data portal have been built based on CKAN.

Data-driven science, or e-Science, is a new paradigm that goes further than mere experimental and theoretical research and computer simulation [6]. In this paradigm, scientific data, which are comprised of observations and the results of scientific activities, are shared and re-used so that scientists can accelerate research activities. Free and open access to publications and the scientific data provided by publicly funded research offers significant social benefits. This has sparked an explosion in the availability of scientific datasets [7], including the raw data obtained by observation and the data derived from computational models and simulations [8].

Such scientific open data and their repositories, including the World Data System (WDS) [*4], Pangaea [*5], and ICPSR, are also described as well as the characteristics of scientific data, especially their spatio-temporal information.

2.2 Searching and Analyzing Scientific Datasets

A large volume of published scientific datasets can be stored on-line in public repositories and made accessible to users within (or without) a scientific community to foster interorganizational and inter-disciplinary research that can accelerate scientific discovery [9], [10]. Such published datasets, which number in the millions, continue to grow impressively and are long-term archived in affordable cloud storage and on disks [11]. Several scientific repositories such as WDS and Pangaea have their own search systems that are designed for discovering scientific datasets. One expected feature for scientific dataset searches is to simultaneously search datasets in several domains. De proposed an search result merging method for metasearch [12].

Many approaches have been proposed to utilize such stored scientific data. For example, we proposed time-series predictions for open scientific data [13]. Fiore created a framework to manage scientific datasets with spatio-temporal information [14]. Steed proposed a web-based climate data analysis framework [15] that visualizes spatio-temporal information or the correlation of target scientific datasets.

Generally, the size of the dataset repository increases monotonically. A distributed database is one solution to keep adequate storage. Xiang proposed the optimization of query for distributed scientific database [16].

2.3 Improving Search Result Performance

The diversity of the search results is one of the most important features for search systems. Especially, trade off between relevance and diversity of search results is one of the issue for search result diversification, and there are several studies to solve it.

One of the approach is to focus on the input of the search procedure, i.e. input query. Hoque proposed a query expansion method considering both diversity and precision of

[*2] http://www.data.gov/
[*3] http://ckan.org/
[*4] http://www.icsu-wds.org/
[*5] http://www.pangaea.de/

the search results for image retrieval [17]. Understanding the query intention is one of the key element of search result diversification because the system can switch the behavior to adjust it. For that purpose, the method proposed by Umemoto predicts the query reformulation type form user action [18]. Hiroshima proposed a concept-based query type inference method [19].

Another approach focus on the output of the search procedure, i.e. the ranking of search results. There are many kinds of studies about ranking or reranking search results. For example, Yan considers the diversity of search results based on latent dirichlet allocation for biomedical document retrieval [20]. Meng ranks search results avoiding duplication of results by considering the relevance and semantics for image retrieval [21]. Reranking algorithm proposed by Zhu tries to avoid redundant search results based on random walks in an absorbing markov chain [22]. Reranking algorithm proposed by Tian considers hierarchical structure of the concept of search targets for image retrieval [23].

On the other hand, it is possible to apply users' feedback to improve the search performance. Relevance Feedback is one of the standard way to improve the precision of search result. In relevance feedback, the users selects some of the search results as query matched results. The system learns the characteristics of these selected results and reflect to next search. Calumby applied genetic programming technique in learning phase of relevance feedback for image search [24]. The search result interface built by Yamamoto enables users to rank by the features of the search targets [25]. Users can select any kinds of features shown in initial search results then search results are reranked by the feature such as keyword or numerical parameters. Xu proposed cluster-based query expansion based on pseudo relevance feedback in biomedical domain [26]. When a search query is too general, search systems return a huge amount of search results. For such case, Battle proposed a result reduction method based on interactive visualization [27]. When the users accept additional manipulations, to apply user's reaction is a good way to improve the search performance.

The spatial information is one of the characteristic information of scientific data and it is useful to increase the diversity of the search results. To the best of knowledge, currently there are no search systems which consider both scientific data search and spatial diversity of the search results.

3. Cross-DB Search System

In this section, we describe the Cross-DB Search System, which is a multi-domain open scientific dataset search system. About 0.8 million pieces of metadata of open scientific data from different domains are indexed in the Cross-DB Search System and can be searched for using a combination of spatial, temporal, and text search queries. The input search query is expanded using query expansion with spatio-temporal information, which is used to visualize the search results. Users can find not only query-matched datasets but also query-related datasets using the Cross-DB Search System.

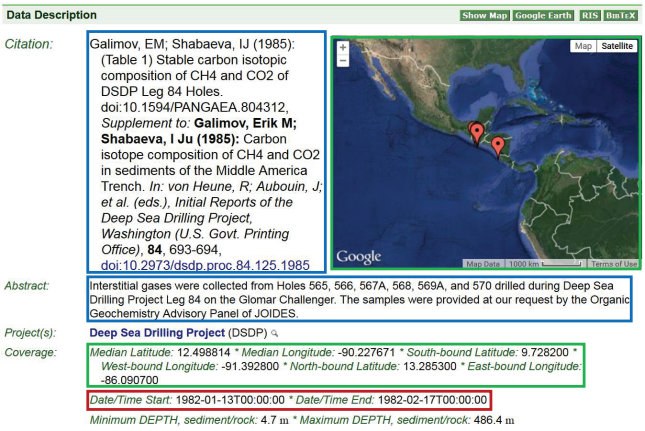


Fig. 2 Example of dataset description that contains spatial, temporal, and several pieces of text information such as title, author, abstract, etc. This information is the dataset's metadata. The variety of information in the metadata is dependent on the dataset domain and the data repository.

3.1 Target Repositories

To build a multi domain dataset search system, scientific datasets from many kinds of repositories are needed to construct database index. Some repositories have standard data providing method. For example, Pangaea provides the metadata of their scientific open data using Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) which is a standard API for metadata harvesting. As previously described in Section 2, CKAN provides several ways to harvest its raw data and the metadata of datasets. However, many other repositories only provide their data on web pages. In this case, we have to harvest their data by parsing HTML and individually extracting the needed information. For this purpose, we built a data harvesting system that aggregates the related web pages of datasets from the repository and extracts the information needed for searching. Note that problems exist with the granularity and the quality of the datasets [28] and copyrights. We only harvested a minimal set of metadata (e.g. spatio-temporal information, text information, author's name, dataset's URI) instead of its raw data.

An index of the Cross-DB Search System is created using the spatial, temporal, and text (STT) information of the harvested datasets. During this process, each dataset is related to the concept of ontologies based on their text information. The concepts of SWEET Ontology [*6] are used for the current Cross-DB Search System. The information of the relationship between datasets and concepts shows the distribution of datasets in ontological networks and similarity calculation described in Section 3.3.

About 0.80 million datasets are harvested from 61 scientific data repositories. They are mainly belonging to WDS.

[*6] http://sweet.jpl.nasa.gov/ontology/

Table 1 Number of scientific data repositories belonging to World Data System and their datasets.

Scientific domain	#Repositories	#Datasets
Earth science	30	492,997
Social science	10	115,875
Biology	6	13,656
Astronomy	3	10,833
Chemistry	2	4,309
Medicine	2	396
others	8	129,349
total	62	807,415

Table 1 shows the number of harvested repositories and datasets for each domain.

3.2 Characteristics of Scientific Datasets

A dataset's various information is treated as metadata. For example, title, creator/publisher's name, its spatio-temporal information, and the name of the obtained parameters are the standard contents of metadata. Since spatio-temporal information is especially useful for dataset search, this section describes the detailed characteristics of a dataset's spatio-temporal information.

Figure 2 shows an example of a dataset description. This dataset has spatial, temporal, and text information such as title, author, and abstract. This information is used as its metadata. The variety of information in the metadata of a dataset depends on its domain and data repository.

A dataset's spatial and temporal information are generally given as a 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 the beginning of 1990 stops at the end of 2000 have $(\boldsymbol{x}_b, \boldsymbol{x_e})$ set as $(1990, 2000)$. For the spatial information, the southernmost point and northernmost 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.

Here we consider the existence ratio of each information type. Table 2 shows the information ratio of the datasets in Pangaea. Although all of datasets have text information such as their own title and author's name, it is not enough amount for text-based search systems. For that purpose, abstract of dataset is important because they contain much text information. Therefore the abstract existence ratio is one of the important features for the systems. However only 1.7% of datasets have abstracts As shown in the table. On the other hand, 73.2% have both space and time information. It can be possible to improve the search performance by applying spatial and temporal information of datasets.

3.3 Dataset Search Process

In this section, we describe the search process of Cross-DB Search System. To solve the problem occurred by the lack of text information, Cross-DB generates STT query from input

Table 2 Information's existence ratio of datasets in Pangaea.

	#Datasets	Ratio
overall	405,456	
w/ abstract	7,028	0.017
w/ time info.	297,478	0.733
w/ space info.	404,145	0.996
w/ space and time info.	297,037	0.732

query so that user can find more spatio-temporally related datasets. The search process is conducted in the following steps.

First, the user inputs keywords, the spatio-temporal conditions, or both, as an input query. A query is composed of a combination of spatial, temporal, and text information and is designated as an STT query. If the input STT query is given only by keywords (text information), then 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. As described 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 a standard technique to represent the similarity between two documents. In this step, the spatio-temporal conditions in an STT query are used simply to find datasets that conform to the given conditions.

For searching through datasets, finding more datasets is crucial, especially those that do not exactly match the input query but that are closely related to it. For that purpose, many search systems use query expansion methods to find more results. We previously proposed a spatiotemporal query expansion method named STT Pseudo Relevance Feedback (STT-PRF) [3]. For standard Pseudo Relevance Feedback (PRF), the algorithm treats the top L datasets in the initial ranking, designated as Y_L, as relevant datasets. Then additional text queries are built from the text information. In STT-PRF, however, the query is composed not solely of text information but also of spatial and temporal information. The beginning and end dates of each dataset in Y_L form time queries; their spatial coverage includes space queries. The set of text, time, and space queries is treated as an expanded query and is used by the second dataset search process.

In the second dataset search process, the space and time scores are calculated for each dataset in addition to the text score. Space score ϕ_s for dataset y is given by the following equations:

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

Here y' shows the dataset in Y_L, which is a set of the datasets treated as relevant ones. d_s stands for the space distances between the two datasets. The time score ϕ_t is also evaluated just like Eq.(1) but using temporal distance d_t.

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

$$\phi(y) = w_s\phi_s(y) + w_t\phi_t(y) + \phi_k(y). \quad (2)$$

Here w_s and w_t are the weight parameters for each distance. After the score calculation, the indexed datasets are ranked by their total scores. The second search process outputs the ranked datasets. This score is called STT score because it uses all of the spatial, temporal, and text scores.

Although PRF uses text information for the second dataset search process, STT-PRF additionally uses space and time information and calculates the distance between two datasets to obtain space and time scores. Several definitions of the spatiotemporal distance (or similarity) exist [29]. These distances are used to calculate $\phi_s(y)$ and $\phi_t(y)$.

Here we describe these distances in detail using information based on the metadata of datasets. First, we define the spatial and temporal distance among the datasets. The spatial and temporal information of a dataset is given as a one- or two-dimensional range. Therefore, we approximate this information using a normal distribution with the following mean μ and variance Σ:

$$\boldsymbol{\mu} = \frac{1}{2}(\boldsymbol{x}_e + \boldsymbol{x}_b), \quad \Sigma = \frac{1}{12}(\boldsymbol{x}_e - \boldsymbol{x}_b)^2. \tag{3}$$

Here $\boldsymbol{x}_b$ and $\boldsymbol{x}_e$ respectively stand for the beginning and end points of the spatial/temporal information. These values are the same as the mean and variance of a uniform distributions with identical $\boldsymbol{x}_b$ and $\boldsymbol{x}_e$.

The distance between datasets is defined using Bhattacharyya distance d_B [30], which is a standard definition for measuring the distance between two probability distributions:

$$d_B(p, p') = -\ln\left(\int \sqrt{p(x)p'(x)}dx\right), \tag{4}$$

Here p and p' are probability distributions that approximate the spatial/temporal information of a dataset. Bhattacharyya distance d_B for normal distributions p_i and p_j is further transformed as follows:

$$d_B(p_i, p_j) = \frac{1}{8}(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)^\top \left[\frac{1}{2}(\Sigma_i + \Sigma_j)\right]^{-1} (\boldsymbol{\mu}_i - \boldsymbol{\mu}_j) + \frac{1}{2}\ln\left\{\frac{\det(\frac{1}{2}(\Sigma_i + \Sigma_j))}{\sqrt{\det(\Sigma_i)\det(\Sigma_j)}}\right\}. \tag{5}$$

4. Reranking Based on Spatial Diversity

4.1 Problem Settings

The problem of placing several objects in two-dimensional map with or without overlap is known as the label placement problem. The target of general label placement problems is to place square labels near the target points, lines, or areas based on positional relations. For example, if the target is a point, the following four points are the candidates of the label places: the upper, lower, left, and right side.

Here we define the this paper' challenge as a rectangular label placement problem that does not allow the overlap of regions. This differentiates datasets from the original search results with spatial constraint. Here the label placement's target is the spatial region of the scientific datasets and the label's position and size equals that of the datasets.

4.2 Spatially Constrained Greedy Selection Method

Next we describe our method to select datasets from the original search results under spatial constraints. The spatial regions of the selected results are used for displaying their spatial regions without any overlap. The proposed scheme, which is named the Spatially Constrained Greedy Selection (SCGS) method, consists of by two steps. Algorithm 1 shows its pseudo code.

Algorithm 1 SCGS method

Input: $\boldsymbol{R}$, $\boldsymbol{S}$
/* Step 1 */
for $i = 1$ to N **do**
  $p \leftarrow 0$
  for $j = i + 1$ to N **do**
    if r_j is occluded by $O(r_i)$ **then**
      $p \leftarrow p + s_j$
    end if
    if $p \geq T$ **then**
      break this loop
    end if
  end for
  if $p < T$ **then**
    $t \leftarrow i$
    break this loop
  end if
end for
/* Step 2 */
for all r_i such that $t \leq i \leq N$ **do**
  add r_i to Y
  for all r_j such that $i + 1 \leq j \leq N$ **do**
    if r_j is occluded by $O(r_i)$ **then**
      pop r_j from $\boldsymbol{R}$
    end if
  end for
end for
Output: Y

First, $\boldsymbol{R}$ is a list of the spatial region of datasets r that are given by the search result and sorted by the value of the STT scores described in Section 3.3. s_i shows the STT score of r_i, and $\boldsymbol{S}$ shows the set of all s_i. Using r_1, which is a spatial region of a dataset with maximums score, remaining spatial regions r_j are checked to determine whether they are overlapped by r_1. If r_j is overlapped by r_1, score s_i is added to penalty p. This check is repeated until p exceeds threshold T. When it becomes $p > T$, the process for r_1 is stopped, and r_1 is removed from the candidate of the display regions. The same process is applied to the second dataset, r_2, and the remaining datasets. When p finally becomes smaller than T, we consider r_i, which is the target of the above process, the result of Step 1.

At the beginning of Step 2, r_i, which is given by Step 1, is added to set Y that contains the datasets for displaying their regions. Next, other regions r_j, which have smaller scores, are checked to see whether they are overlapped by r_i. Any r_j overlapped by r_i is removed from the display candidate. After every r_j is checked, this process is repeated with

Table 3 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	

r that has the next largest score. This process is repeated until every r is removed as a display candidate or added to Y.

To increase the diversity of search results, the selected non-overlapped datasets should obtain higher rank. In this paper, we simply rerank these datasets to top of the search results sorting with STT score. As a result, we obtain the reranked search results whose spatial regions of top results are not overlapped each other.

5. Performance Evaluation

5.1 Experimental Setup

In this section we describe the datasets for performance evaluation. Since the performance of a keyword-based search is largely dependent on the ability of users to formulate good queries, it must be evaluated with commonly used queries. We used keywords from actual query lists obtained from major search engines that were chosen from actual keywords related to science and obtained from Google Trends [*7] and the query logs of the Cross-DB search system. Additional keywords were chosen from environmental science fields by Microsoft Academic Search [*8] for the current trends in the searched for terms. More keywords were chosen from ontological concepts and created using the SWEET ontology, which mainly covers the earth and environmental science terms. These keywords, which were selected from natural science domains, are presented in Tab. 3.

For our experiments, we took datasets from the Pangaea database. For each test query, we collected the top 120 ranked datasets identified by Pangaea's search system and used them as queries. The relevance of all the retrieved datasets was manually evaluated by three human labelers with master's degrees in natural science. According to the queries, the relevance of the retrieved datasets was 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. In the following experiments, datasets with relevance values of 2 or 3 were considered *query-related*. These testsets were designed to evaluate scientific search systems by all those who want to evaluate their methods by accessing and using such systems [*9].

For our experiments, since the amount of available data was limited, we empirically determined the values of weight parameters w_s and w_t in Eq. (2) to be 0.370 and 0.074. The parameter L in Eq. (1) was set to 10. The threshold T in algorithm 1 is set to 10 which is given by the average number of overlapped datasets.

5.2 Experimental Results

To evaluate the SCGS method, we used the following four comparison methods.

- AKMY: As one schema to solve the label placement method, we use the AKMY method proposed by Abe et al [31]. This method was originally designed to solve n-point label placement problems. Since the experiments in this section are 1-point label placement problems, we arranged and applied the method. We set the size of the labels to the same size as the target region of the datasets.
- STT: We only executed Step 2 of the algorithm 1. The difference between the following Random method is that $\boldsymbol{R}$ is sorted by STT scores.
- Random: In algorithm 1, $\boldsymbol{R}$ is not sorted by scores but is listed randomly. Only Step 2 of the algorithm is executed.
- Exclusive: This method only select the regions that are not overlapped by any other regions.

Note that SCGS, AKMY and STT method use STT scores in their algorithm. Random and Exclusive method are used to show the effectiveness of using STT score.

We consider that the total score of datasets which are not overlapped each other represents the quality of the reranked search results. Non-overlapped datasets obtain higher rank by the proposed method so that reranked results have both spatial diversity and relevancy to the query. At this point of view, we use the total score of non-overlapped datasets in the search results as the evaluation criteria. We also measure the number of non-overlapped datasets as another criteria. Note that the region coverage in whole map is not a suitable criteria because there are some datasets which spatially cover all over the world. In that case, only one datasets can bring high coverage.

Table 4 shows the total score of datasets selected by each methods when we rerank top n datasets in search results. This result shows that the SCGS outperforms other methods. The result of SCGS is already converged at result of 10. The reason is that significant percentage of datasets in original search results have same spatial information so that most of them are not selected. Note that the number of datasets are not monotonically increase because they does not select datasets cumulatively.

The quality of the first result page is the most important

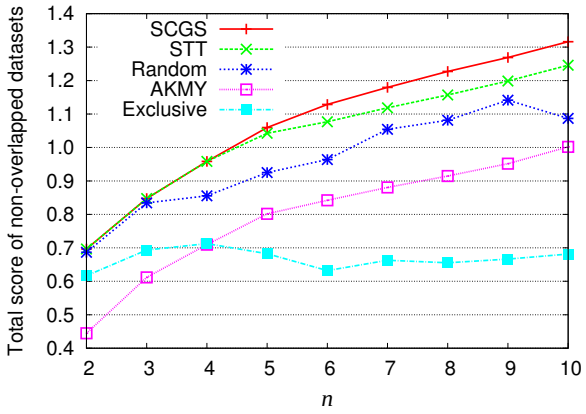
[*7] http://www.google.com/trends/
[*8] http://academic.research.microsoft.com/
[*9] This evaluation dataset is available at http://www2.nict.go.jp/univ-com/isp/s.takeuchi/sttprf.tgz

Table 4 Total score of top n datasets in reranked search results

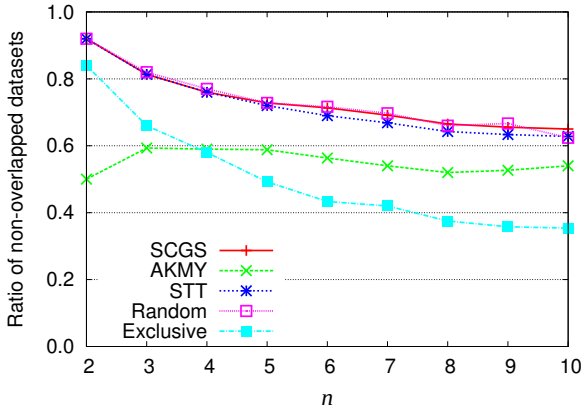
n	SCGS	AKMY	STT	Random	Exclusive
5	1.059	0.801	1.042	0.99	0.682
10	1.316	1.001	1.245	1.13	0.681
20	1.316	1.028	1.246	1.08	0.564
50	1.316	0.992	1.246	0.99	0.245
70	1.316	1.001	1.246	0.95	0.181
100	1.316	1.001	1.246	1.06	0.181
120	1.316	0.980	1.246	0.93	0.181

for search systems. Generally, they shows 10 to 20 results in the first result page. Therefore we focus on the top 10 datasets of the search results for the detailed comparison. Figure 3 shows (a) total score and (b) ratio of reranked datasets in top n rank in reranked search results. As shown in these Fig. 3 (a), SCGS outperforms other methods in total score independent of n. Note that the result of "Exclusive" does not monotonically increase because they does not select datasets cumulatively. Although the Fig. 3 (b) does not show quite difference between SCGS, STT, and Random, SCGS slightly outperforms them.

Figure 4 shows examples of selected regions for the search result of "environmental impact" by (a) SCGS method, (b) AKMY, (c) STT, and (d) Random result. In each figure, the region of selected and reranked regions are indicated by rectangles. SCGS provides best results in the number of datasets and total score.



(a) total score of non-overlapped datasets



(b) Ratio of non-overlapped datasets

Fig. 3 (a) total score and (b) ratio of reranked datasets in top n rank in reranked search results. As shown in (a), SCGS outperforms other methods in total score. Note that the result of "Exclusive" does not monotonically increase because they does not select datasets cumulatively. Although (b) does not show quite difference between SCGS, STT, and Random, SCGS slightly outperforms them.

6. Conclusion

In this paper, we propose a novel rectangular label placement method to rerank search results based on spatial diversity for scientific data search systems. Selecting the spatial regions of datasets, given as the results of scientific dataset search, can be treated as a label placement problem that allows label removal, which is considered NP-hard. Our proposed Spatially Constrained Greedy Selection (SCGS) method selects datasets with highly query-related datasets. Since these datasets without overlap brings spatial diversity of search results, and users can easily overview the search results. We describe the relevance of the datasets but they are not applied in this paper. As the future work, the relevance of the datasets will be used as one of the independent rationale.

References

[1] JETRO: Open data. available from ⟨http://www.ipa.go.jp/files/000033718.pdf⟩ (2013).

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

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

[4] Hendler, J., Holm, J., Musialek, C. and Thomas, G.: US Government Linked Open Data: Semantic.data.gov, *Intelligent Systems, IEEE*, Vol. 27, No. 3, pp. 25–31 (2012).

[5] Hoxha, J. and Brahaj, A.: Open Government Data on the Web: A Semantic Approach, *Emerging Intelligent Data and Web Technologies (EIDWT), 2011 International Conference on*, pp. 107–113 (2011).

[6] Hey, T., Tansley, S. and Tolle, K.: The Fourth Paradigm: Data-Intensive Scientific Discovery, *Microsoft Research* (2009).

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

[8] Yu, J. and Buyya, R.: A Taxonomy of Workflow Management Systems for Grid Computing, *ACM SIGMOD Record*, Vol. 34, No. 3, pp. 44–49 (2005).

[9] Humphrey, M., Agarwal, D. and van Ingen, C.: Fluxdata.org: Publication and Curation of Shared Scientific Climate and Earth Sciences Data, *5th IEEE International Conference on e-Science(e-science 2009)*, pp. 118–125 (2009).

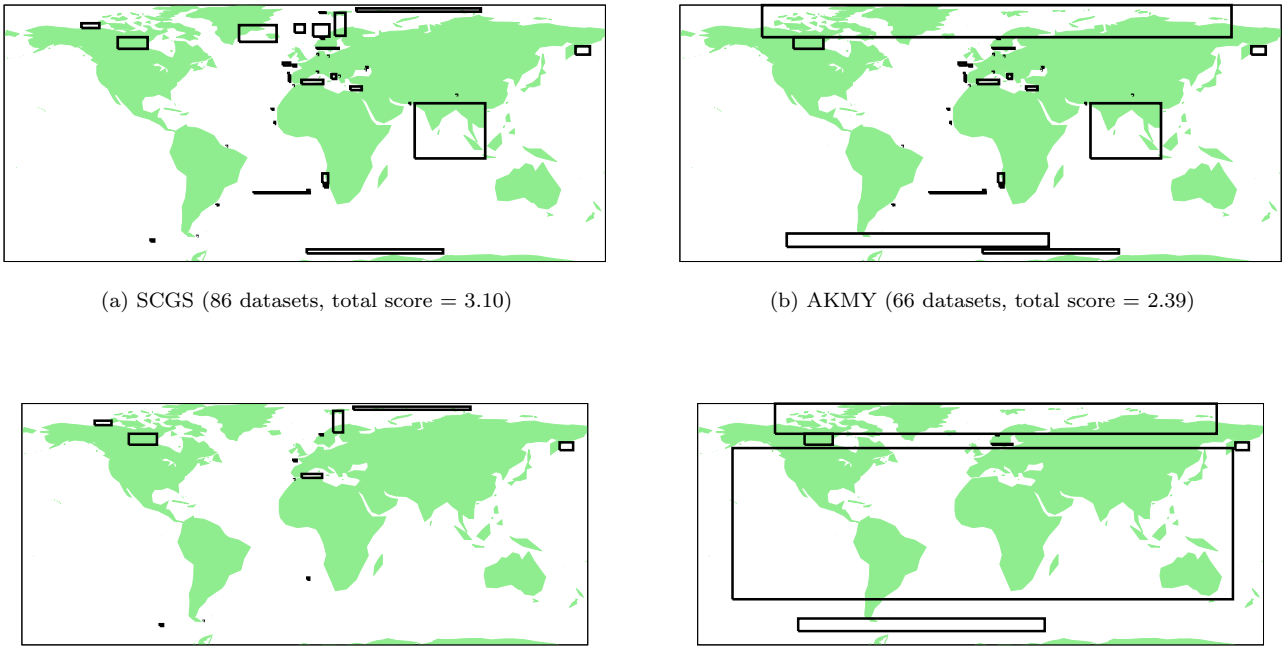
[10] on Issues in the Transborder Flow of Scientific Data, C.: *Bits of Power: Issues in Global Access ot Scientific Data*, National Academy Press (1997).

[11] Gray, J., Szalay, A. S., Thakar, A. R., Stoughton, C. and Vandenberg, J.: Online Scientific Data Curation, Publication, and Archiving, *SPIE Astronomy Telescopes and Instruments*, No. MSR-TR-2002-74, p. 6 (2002).

[12] De, A., Diaz, E. and Raghavan, V.: Search Engine Result Aggregation Using Analytical Hierarchy Process, *Web Intelligence and Intelligent Agent Technology (WI-IAT), 2010 IEEE/WIC/ACM International Conference on*, Vol. 3, pp. 300–303 (2010).

[13] Ong, B., Sugiura, K. and Zettsu, K.: Dynamic pre-training of Deep Recurrent Neural Networks for predicting environmental monitoring data, *Big Data (Big Data), 2014 IEEE International Conference on*, pp. 760–765 (2014).

[14] Fiore, S., Palazzo, C., D'Anca, A., Foster, I., Williams, D. and Aloisio, G.: A big data analytics framework for scientific data management, *Big Data, 2013 IEEE International*



(a) SCGS (86 datasets, total score = 3.10)

(b) AKMY (66 datasets, total score = 2.39)

(c) STT (23 datasets, total score = 0.63)

(d) Random (14 datasets, total score = 1.09)

Fig. 4 Examples of selected regions for search results of "environmental impact": (a) SCGS method, (b) AKMY, (c) STT, and (d) Random result. In each figure, the region of selected and reranked datasets by each methods are indicated by rectangles. SCGS provides best results in number of datasets and total score.

Conference on, pp. 1–8 (2013).

[15] Steed, C., Evans, K., Harney, J., Jewell, B., Shipman, G., Smith, B., Thornton, P. and Williams, D.: Web-based visual analytics for extreme scale climate science, *Big Data (Big Data), 2014 IEEE International Conference on*, pp. 383–392 (2014).

[16] Xiang, H.: Query optimization over a heterogeneously distributed scientific database, *Big Data, 2013 IEEE International Conference on*, pp. 58–64 (2013).

[17] Hoque, E., Hoeber, O. and Gong, M.: Evaluating the Trade-Offs between Diversity and Precision for Web Image Search Using Concept-Based Query Expansion, *Web Intelligence and Intelligent Agent Technology (WI-IAT), 2011 IEEE/WIC/ACM International Conference on*, Vol. 3, pp. 130–133 (2011).

[18] Umemoto, K., Nakamura, S., Takehiro, Y. and Tanaka, K.: Predicting Query Reformulation Type from Web Searcher Behavior, *IPSJ Transactions on databases*, Vol. 6, No. 3, pp. 132–147 (2013).

[19] Hiroshima, N., Toda, H., Matsuura, Y. and Kataoka, R.: A Query Type Inference Method Using Concept Base and Its Evaluation, *IPSJ Transactions on databases*, Vol. 3, No. 3, pp. 33–45 (2010).

[20] Chen, Y., Yin, X., Li, Z., Hu, X. and Huang, J.: Promoting Ranking Diversity for Biomedical Information Retrieval Based on LDA, *Bioinformatics and Biomedicine (BIBM), 2011 IEEE International Conference on*, pp. 456–461 (2011).

[21] Wang, M., Yang, K., Hua, X.-S. and Zhang, H.-J.: Towards a Relevant and Diverse Search of Social Images, *Multimedia, IEEE Transactions on*, Vol. 12, No. 8, pp. 829–842 (2010).

[22] Zhu, X., Goldberg, A. B., Van Gael, J. and Andrzejewski, D.: Improving Diversity in Ranking using Absorbing Random Walks., *HLT-NAACL*, Citeseer, pp. 97–104 (2007).

[23] Tian, X., Yang, L., Lu, Y., Tian, Q. and Tao, D.: Image Search Reranking With Hierarchical Topic Awareness, *Cybernetics, IEEE Transactions on*, Vol. PP, No. 99, p. 1 (2015).

[24] Calumby, R., da Silva Torres, R. and Goncalves, M.: Diversity-driven learning for multimodal image retrieval with relevance feedback, *Image Processing (ICIP), 2014 IEEE International Conference on*, pp. 2197–2201 (2014).

[25] Yamamoto, T., Satoshi, N. and Tanaka, K.: RerankEverything: A Reranking Interface for Browsing Ranked Results Flexibly, *IPSJ Transactions on databases*, Vol. 3, No. 4, pp. 48–64 (2010).

[26] Xu, X. and Hu, X.: Cluster-based query expansion using language modeling in the biomedical domain, *Bioinformatics and Biomedicine Workshops (BIBMW), 2010 IEEE International Conference on*, pp. 185–188 (2010).

[27] Battle, L., Stonebraker, M. and Chang, R.: Dynamic reduction of query result sets for interactive visualization, *Big Data, 2013 IEEE International Conference on*, pp. 1–8 (2013).

[28] Knap, T., Michelfeit, J. and Necasky, M.: Linked Open Data Aggregation: Conflict Resolution and Aggregate Quality, *Computer Software and Applications Conference Workshops (COMPSACW), 2012 IEEE 36th Annual*, pp. 106–111 (2012).

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

[30] Bhattacharyya, A.: 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).

[31] Abe, N., Amai, Y., Nakatake, T., Masuda, S. and Yamaguchi, K.: An Algorithm for the Map Labeling Problem with Two Kinds of Priorities, *International Journal of Mathematical, Computational, Statistical, Natural and Physical Engineering*, Vol. 8, No. 5, pp. 800 – 803 (2014).