Using Multiple Viewpoints to Improve Access to Earth Science Data

James C. French A. C. Chapin Worthy N. Martin
Department of Computer Science
University of Virginia
Charlottesville, VA
{french|acc2a|wnm}@cs.virginia.edu

Abstract— **To retrieve information from a collection of data objects a system must specify a representational scheme and a corresponding mechanism for accessing content. Such representations and access mechanisms are usually discipline-specific and here we call them viewpoints. We are investigating the advantages of making use of multiple viewpoints in a single retrieval system. Consider the computer science literature indexed in the ACM Digital Library using the Computing Classification System, and also in INSPEC under the INSPEC subject headings; each of these classification systems uses its own controlled vocabulary and provides a different viewpoint on the same data.**

Access to Earth science data via the EOS Data Gateway (EDG) or the Global Change Master Directory (GCMD) is provided by a variety of strategies. Free-text searching is provided (a keyword viewpoint) and the data is also accessible by "valids" (a subject heading viewpoint). In addition there are spatial and temporal viewpoints. In this paper we discuss our strategy for combining these different viewpoints into a cohesive search strategy that we hope will increase the searcher's ability to locate relevant data. This requires a framework in which a searcher can pose queries in one viewpoint and then change to another viewpoint while retaining a sense of context. We will describe our framework and demonstrate its utility with a concrete example.

I. Introduction

THE search for a particular information need within a document collection typically begins with the submission of a description of the information need – a query – to an information retrieval system that can access descriptions of the documents in the collection and make judgments about which document descriptions represent the best responses to the query.

Formulating a query effectively depends on consistent use of vocabulary between the query and the document descriptions. Searchers new to a field of interest may be stymied by unfamiliarity with the fields shorthands and shibboleths, or confused by conflicting usage of common terms in different fields.

This work supported in part by NASA Grants NAG5-8585 and NAG5-9747 and NASA GSRP NGT5-50062.

For example, in the NASA earth science data store, datasets from many earth science disciplines are collected in one place and indexed based on a controlled vocabulary called valids. Aside from the difficulty of coming to an agreement on the use of any such controlled vocabulary, and enforcing its consistent use in describing the datasets, this approach means that a searcher must become familiar with the controlled vocabulary before being able to search this earth science data store effectively. In fact, a novice searcher is likely to conclude that the data store contains no information on a particular topic of interest when in fact, there may be a trove of such information, described by some obscure valid unknown to the searcher.

If there existed some mapping between the searcher's vocabulary and the valids then the searcher's goal could be found without the searcher having to learn the appropriate valids. This is one use of multiple viewpoint systems

A viewpoint is a representational scheme on a collection of data objects, and a corresponding mechanism for accessing the data. Retrieving datasets keyed on valids constitutes one viewpoint. If we add a second viewpoint using a different vocabulary, we expand the set of queries which can produce responses. Mapping from one viewpoint to another can not only help the user maintain a sense of context when searching in different vocabularies but can also improve search results. A "thesaurus viewpoint" that represents relationships between words in different vocabularies can extend the set of productive queries without the new vocabulary being directly linked to the underlying data collection. We also believe that viewpoints can be a useful tool for information exploration and browsing activities.

As one example, the web sites of the ACM Digital Library[1], IEEE Digital Library[2], and INSPEC[3] index

[1] http://www.acm.org/dl/
[2] http://www.computer.org/publications/dlib
[3] http://www.ieee.org.uk/publish/inspec/

of publications constitute an ad hoc multiple viewpoint system on technical papers.

Suppose a searcher submits the query "quad trees" to all these sites, and finds that the ACMDL gives the largest set of relevant papers, while the IEEEDL gives relatively few. A searcher experienced in using these sites in concert would not then assume the IEEEDL lacked useful information but would attempt to approach it in a new way, for instance using an author from the ACMDL list as a query to the IEEEDL and thereby discovering that the query is better phrased in IEEE parlance as "nearest neighbor methods." A well-designed multiple viewpoint system incorporating these web sites could do such a transition on author name automatically to help a novice searcher overcome such vocabulary mismatches.

II. Context

The term "viewpoints" has been used with some frequency, and great inconsistency, in the areas of information visualization and user interfaces. Researchers such as Teraoka and Maruyama [1] are usually interested in representing a searcher's interests and purpose, and "multiple viewpoints" in this case consist of parameters to an information visualization system which indicate how to present information based on a particular (dynamically evolving) interest profile. While a viewpoint in our parlance might well be based on a particular user's interests, we do not limit differences among viewpoints to different visualizations of the same information relationships; in our viewpoints, the relationships may differ as well.

Wilbur [2] found that using the relevance judgments of several persons in retrieval systems improves on retrieval rankings in comparison to using the relevance judgment of a single individual, and Rajashekar and Croft [3] report consistent improvement in document retrieval when two or more document representations were combined, over using a single representation. The concept of multiple viewpoints as we use it was introduced by Powell and French [4] as an approach to taking advantage of these findings. Multiple viewpoints allow both different relevance judgements and different representations to be used together in a single information retrieval system.

Buckland's Unfamiliar Metadata project [5] approaches the problem of mismatched vocabulary between searcher and index by suggesting query augmentation. An EVI (Entry Vocabulary Index) is a mapping from an ordinary language query to a list of possibilities drawn from the vocabulary used in the database. This is an example of using two viewpoints, the database itself indexed by its own vocabulary, and an "EVI viewpoint" which represents relationships between words in the two vocabularies, and whose output can be used as input to the other. A study demonstrating the potential of this approach can be found in French *et al.*[6].

III. Defining Viewpoints

In describing a system of multiple viewpoints, we first identify the universe of data items with which we are concerned. This may include not only artifacts from a collection, such as the items in a database or books in a library, but also auxiliary data items such as keywords and saved queries.

Several data decompositions can be applied to the universe of data items, organizing it into categories; separating keywords from books is one such decomposition, but books might also be decomposed into fiction and nonfiction. Data decompositions are suitable for broad, rigid categorizations, but lack the fluidity of description available with viewpoints. A lens is a process which intersects some set defined in a data decomposition with another subset of the universe of data items; for instance, a lens for nonfiction books could be applied to the result of a library search so that only nonfiction books satisfying the search are shown.

Each viewpoint provides a representational scheme for some subset of the universe of data items, its viewpoint data set, and a mechanism for accessing this content. The set of possible inputs to the access mechanism are viewpoint queries. Each query has a corresponding viewpoint result, which is some subset of the viewpoint data set.

A text query to a multiple viewpoint system might have to undergo some translation and augmentation before it is usable as a viewpoint query for a particular viewpoint; the initial query transition performs this transformation. For each pair of viewpoints there is a transition mapping which, given a query in the first viewpoint, produces a query in the second viewpoint.

A result merge function takes all the viewpoint results that have been created from querying system viewpoints and organizes them into a system result suitable for displaying to the user. This may be done dynamically during the search process, showing the user how the result changes with different queries and different viewpoints, and may include or not include elements of any viewpoint result. Typically the universe of data items will be decomposed to discriminate appropriate result data from keywords, et cetera, and the corresponding lens will be applied to the system result.

A. Interaction Styles

Interaction with a system of multiple viewpoints may take place at several levels. The structure of separate viewpoints may be hidden from searchers, with any transitions among viewpoints, query transformations, and result merging occurring silently, so that it will appear that the user is dealing with a single viewpoint which maps an initial query to a final result. Metasearch-engines on the web, which send their queries to other search engines and then collate the results are an example of a system of multiple viewpoints operating in this way. Designing a search strategy to take advantage of multiple viewpoints without intermediate user interaction is an interesting research area.

At the other extreme, the user might be offered the ability to move among viewpoints, change queries, and alter merging strategies with little input from the system. This approach is probably best suited to information exploration and browsing tasks. Maintaining both context and a sense of purpose in such an interaction is a major HCI research problem.

IV. Our Approach and An Example

NASA datasets are stored at Distributed Active Archive Centers (DAAC). These datasets record observations from projects and store these observations as granules which are data files and are the means of distributing the data to users. The granules often differentiate the data along spatial and/or temporal dimensions.

The datasets are described by subject heading metadata called *valids*. The valids are a controlled vocabulary and are used for searching within EOSDIS. We have also extracted another set of dataset descriptions based on text passages.

So, the universe of data items for our multiple viewpoint system here consists of the data sets, the valids, and the free text terms, by which categories these data items are decomposed.

A. Concept Spaces

"Concept spaces" have appeared in several guises[7], [8] over the years and are intimately related to our concept of viewpoints

For our purposes, a concept space is an m-dimensional index space induced by a vocabulary of m indexing terms. Each indexed item is represented as a vector in this space. We are using the term "concept space" in preference to "vector space" to differentiate multiple spaces in which we conceptualize the datasets differently.

We are specifically interested in a *free text* space (t-space viewpoint) and a *valids* space (v-space viewpoint). Each of these is a viewpoint in the system. The free text space is derived from descriptive text passages associated with datasets. The valids space is determined by the EOSDIS valids assigned to the datasets. Note that all the objects of interest to us (queries, datasets, and valids) are representable in each viewpoint.

B. Multiple Viewpoints

In earlier work, Powell and French[4] have demonstrated the potential of *multiple viewpoints* to increase retrieval effectiveness by enhancing the discovery process. We have explicitly provided a mechanism in our prototype for switching from a t-space viewpoint search to one in v-space viewpoint to examine the hypothesis that retrieval effectiveness can be improved by searching initially in one viewpoint and then switching to the other via a viewpoint transition mapping.

A third viewpoint, a thesaurus, may also be added to this system, with a lens for valids or for text terms applied to its results; this viewpoint aids in query augmentation for free text queries.

We envisage two modes of interaction with a fully realized multiple viewpoint system for NASA Earth science information systems. A user may simply enter a query by picking valids from a list or by entering a text query. The system will search the v-space viewpoint first in this case, then try the t-space viewpoint if there are non-valid terms in the query; if the result is insufficient (for instance, empty), the user will be prompted with results from the thesaurus viewpoint. The user will also be offered the option of using the thesaurus viewpoint in formulating the initial query.

A more complex user interaction might offer the user explicit access to the different viewpoints, allowing a tight feedback loop for query transformation, and allowing the user to explicitly apply valid or dataset lenses to the results.

The t-space and v-space viewpoints are explored in more concrete terms in the following sections.

B.1 t-space viewpoint

We constructed a text space (t-space viewpoint) by associating descriptive texts with datasets and then using the vector space model (VSM) of information retrieval[9]. In the VSM we represent an object, O_i as a vector, $(w_{i1}, w_{i2}, ..., w_{in})$, in an n-dimensional term space derived from the terms in all the objects. The vector component, w_{ij}, is a weight representing how well term j characterizes object i. We use a $tf \times idf$

weighting strategy where weights have the general form

$$w_{ij} = tf_{ij} \cdot \frac{N}{df_j}.$$

Here tf_{ij} is the *term frequency* of term j, that is, how often term j occurs in object i. The denominator, df_j is called the *document frequency* of term j and denotes the number of objects containing at least one instance of term j.

An example of search results from a t-space viewpoint is shown in Figure 1. We used the dataset summary taken from the DIF[4] entry associated with each dataset to form a text description (representative) for the dataset.

Figure 1 shows the results of processing the query *atmospheric pollution* in the t-space viewpoint representations of all the datasets. The search result is a ranked list of datasets.

Note that each figure shows the valids associated with each ranked dataset. These valids enable a transition from the t-space viewpoint to the v-space viewpoint described in the next section.

B.2 v-space viewpoint

We form the v-space viewpoint by creating a vector, $(v_1, v_2, ..., v_n)$, for each dataset where $v_k = 1$ if valid k is assigned to the dataset. In our prototype we currently use the Jaccard coefficient to measure similarity in the v-space viewpoint.

As stated in the last section, the valids associated with a dataset are used to transition from the t-space viewpoint to the v-space viewpoint. Figure 2 shows an example where the v-space viewpoint has been entered with a focus on the first dataset. We also show the three most similar datasets in the figure.

The following conventions are used in Figure 2. ALL CAPS in the "matched" field of D_k indicates a term that has been assigned to both D_1 and D_k; lowercase indicates a term that is assigned to D_1 and not D_k. The "unmatched" terms are those assigned to D_k and not to D_1.

The user can select any dataset shown and "refocus" attention in the viewpoint to that dataset. In this way it is possible to explore neighborhoods of a dataset for other relevant data.

Note that we can represent individual valids in the t-space viewpoint. By this device we can provide a transition to the t-space viewpoint from the v-space viewpoint. We can also enter the t-space viewpoint via the current dataset under focus in the v-space viewpoint. Both these entry mechanisms into the t-space viewpoint can be used to support a multiple viewpoint interaction.

C. An Example

In this section we give a concrete example to illustrate our viewpoint concepts. We describe a particular search/browse session using our prototype data-space viewpoint browser.[5] We begin with a free text search using the phrase "atmospheric pollution." Figure 1 shows the first 10 datasets returned in response to this query. Note that the acronym `MOPITT` stands for "Measurement of Pollution in the Troposphere." The `MOPITT` and `MAPS` (Measurement of Air Pollution from Satellites) data are reasonable responses to the query.

At this point we note that many of the seemingly relevant datasets have associated valids `CARBON MONOXIDE` and `METHANE`. As a strategy at this point we select the fourth dataset, `MOPO2`, and transition to the "valids viewpoint" with this data set as context.

This results in a viewpoint of the data where we find, among other things, the `ATMOS` (Atmospheric Trace Molecule Spectroscopy Data From Spacelab-3) dataset. According to the `ATMOS` dataset summary: "The Atmospheric Trace Molecule Spectroscopy (ATMOS) experiment was flown on the Space Shuttle STS-51B as part of the Spacelab-3 laboratory to demonstrate the capability to monitor environmental quality by surveying the atmosphere for trace constituents and by identifying their sources, flow patterns, and decay mechanisms." Inspection of the valids associated with the `ATMOS` dataset together with this summary clearly indicates that the `ATMOS` dataset has potential to be relevant to a query on atmospheric pollution.

We now choose to remain in the present viewpoint but change our vantage by recentering on the `ATMOS` dataset. The first four "nearby" datasets are shown in Figure 2. These datasets also clearly have potential relevance to queries on atmospheric pollution.

It is important to note that a direct search of the data using the keywords "carbon monoxide methane" would not have found the ATMOS dataset nor any other dataset shown in Figure 2. To find them we needed a different viewpoint.

V. Future Work

Although in some cases it will be reasonable to create transition mappings from human judgements, in cases where we have sufficient data in two viewpoint representations, we may also attempt to extract some reasonable such mappings using automatic means. Typically, such

[4] Directory Interchange Format

[5] `http://www.cs.virginia.edu/~cyberia/EVOC/DEMO/nasa2/`.

```
Search terms: atmospheric pollution

1: 0.23
dataset: MOPITT Level-3 Data (Gridded CH4 Total Column): MOP05
valids:  METHANE;

2: 0.22
dataset: MOPITT Level-3 Data (Gridded CO Total Column): MOP07
valids:  CARBON MONOXIDE;

3: 0.22
dataset: MOPITT Level-3 Data (Gridded CO Mixing Ratios): MOP06
valids:  CARBON MONOXIDE;

4: 0.22
dataset: MOPITT Level-2 Data from EOS Terra (MOP02)
valids:  CARBON MONOXIDE; METHANE;

5: 0.20
dataset: MOPITT Level-1 Data from EOS Terra (MOP01)
valids:  AEROSOL RADIANCE; OUTGOING LONGWAVE RADIATION; INFRARED IMAGERY;

6: 0.14
dataset: Measurement of Air Pollution from Satellites (MAPS)
         Space Radar Laboratory - 1 (SRL1) Carbon Monoxide
         Second by Second data
valids:  CARBON MONOXIDE;

7: 0.14
dataset: Measurement of Air Pollution from Satellites (MAPS)
         Space Radar Laboratory - 1 (SRL1) Carbon Monoxide
         5 degree by 5 degree data
valids:  CARBON MONOXIDE;

8: 0.11
dataset: Priority Programme for China's Agenda 21
valids:  AGRICULTURAL RESOURCES; ENVIRONMENTAL INDICATORS; INDUSTRIAL
         RESOURCES; AGRICULTURAL EQUIPMENT; FARM STRUCTURES; CROPPING SYSTEMS;
         DAIRY PRODUCTS; LIVESTOCK PRODUCTS; POULTRY PRODUCTS; ANIMAL
         MANAGEMENT SYSTEMS; FIELD CROPS PRODUCTS; FRUIT PRODUCTS;
         HORTICULTURAL PRODUCTS; VEGETABLE PRODUCTS; AGRICULTURAL ECONOMICS;

9: 0.09
dataset: Directory of EuroMAB Biosphere Reserves
valids:  CLIMATE CHANGE; LAND CHARACTERISTICS;

10: 0.09
dataset: Atmospheric Profiles: TOVS - NOAA (FIFE)
valids:  OZONE; ATMOSPHERIC PRESSURE; AIR TEMPERATURE; CLOUD AMOUNT;
```

Fig. 1. Datasets returned for the query "atmospheric pollution." The dataset representations were mined from DIF entries.

mappings will be based on frequency of collocation – how often the same terms are used to describe a data item in two different viewpoints. One approach for finding associations of terms based on frequency that we are exploring is data mining.

VI. Conclusions

Multiple viewpoint systems take advantage of several different sets of comparative judgments about some collection of data in retrieving information from a data collection.

Among the benefits we expect from multiple viewpoint systems are the expansion of available vocabulary for search queries, and the improvements in document retrieval observed in systems using more than one set of judgments about data. We also hope the concept of viewpoints will aid in the thoughtful design of systems to take advantage of these benefits.

References

[1] Terukhio Teraoka and Minoru Maruyama, "Adaptive Information Visualization Based on the User's Multiple Viewpoints – interactive 3D Visualization of the WWW ," in *Proc. 1997 IEEE Symposium on Information Visualization*, 1997.

[2] John W. Wilbur, "The Knowledge in Multiple Human Relevance Judgements," *ACM Transactions on Information Systems*, vol. 16, no. 2, pp. 101–126, 1998.

[3] T. B. Rajashekar and W. Bruce Croft, "Combining Automatic and Manual Index Representation in Probabilistic Retrieval.," *Journal of the American Society for Information Science*, vol. 46, no. 4, 1995.

[4] Allison Powell and James C. French, "The Potential to Improve Retrieval Effectiveness with Multiple Viewpoints," Tech. Rep. CS-98-15, Department of Computer Science, University of Virginia, 1998.

[5] Michael Buckland et al., "Mapping Entry Vocabulary to Unfamiliar Metadata Vocabularies," in *D-Lib Magazine*. January 1999, http://www.dlib.org/dlib/january99/buckland/-01buckland.html.

[6] James French, Allison Powell, Fred Gey, and Natalia Perelman, "Exploiting A Controlled Vocabulary to Improve Collection Selection and Retrieval Effectiveness," in *Proc. Tenth International Conference on Information and Knowledge Management*, 2001, pp. 199–206.

[7] H. Chen, T. D. Ng, J. Martinez, and B. R. Schatz, "A Concept Space Approach to Addressing the Vocabulary Problem in Scientific Information Retrieval: An Experiment on the Worm Community System," *Journal of the American Society for Information Science*, vol. 48, no. 1, pp. 17–31, 1997.

[8] P. Schauble, "Thesaurus Based Concept Spaces," in *Proceedings of the 10th International Conference on Research and Development in Information Retrieval*, New Orleans, LA, 1987, pp. 254–262.

[9] G. Salton and M. J. McGill, *Introduction to Modern Information Retrieval*, McGraw-Hill Book Company, 1983.

```
Atmospheric Trace Molecule Spectroscopy (ATMOS) Data From Spacelab-3

Valids:
  AEROSOL EXTINCTION; CARBON MONOXIDE; NITROGEN OXIDES;
  STRATOPAUSE; CARBON DIOXIDE; CARBONYL SULFIDE; CHLORINE NITRATE;
  CHLOROFLUOROCARBONS; HALOCARBONS; METHANE; NITRIC ACID;
  NITROGEN DIOXIDE; NITROUS OXIDE; OXYGEN; OZONE; TRACE ELEMENTS;
  TRACE GASES; ATMOSPHERIC PRESSURE; AIR TEMPERATURE; WATER VAPOR;
  UPPER LEVEL WINDS; SOLAR RADIATION; REFLECTED INFRARED;
  SOLAR ACTIVE REGIONS
*****
UARS Cryogenic Limb Array Etalon Spectrometer (CLAES) Level 3A
   Data Products via WWW

Valids:
  AEROSOL EXTINCTION; carbon monoxide; nitrogen oxides;
  stratopause; carbon dioxide; carbonyl sulfide; CHLORINE NITRATE;
  CHLOROFLUOROCARBONS; halocarbons; METHANE; NITRIC ACID;
  NITROGEN DIOXIDE; NITROUS OXIDE; oxygen; OZONE; trace elements;
  TRACE GASES; atmospheric pressure; AIR TEMPERATURE; WATER VAPOR;
  upper level winds; solar radiation; reflected infrared;
  solar active regions
Unmatched:
  aerosol radiance; infrared flux
*****
Atmospheric Boundary Layer Experiment (ABLE_3A) Electra Chemical
   Data from the Global Tropospheric Experiment (GTE)

Valids:
  AEROSOL EXTINCTION; CARBON MONOXIDE; NITROGEN OXIDES;
  stratopause; CARBON DIOXIDE; carbonyl sulfide; chlorine nitrate;
  chlorofluorocarbons; halocarbons; METHANE; NITRIC ACID;
  nitrogen dioxide; nitrous oxide; oxygen; ozone; trace elements;
  TRACE GASES; ATMOSPHERIC PRESSURE; AIR TEMPERATURE; water vapor;
  upper level winds; SOLAR RADIATION; reflected infrared;
  solar active regions
Unmatched:
  aerosol radiance; barometric altitude; nitrate particles;
  non-methane hydrocarbons; sulfate particles; tropospheric ozone;
  ultraviolet flux; ultraviolet sensor temperature
*****
UARS Halogen Occultation Experiment (HALOE) Level 3A Data
   Products via WWW

Valids:
  AEROSOL EXTINCTION; carbon monoxide; NITROGEN OXIDES;
  stratopause; carbon dioxide; carbonyl sulfide; chlorine nitrate;
  chlorofluorocarbons; halocarbons; METHANE; nitric acid;
  NITROGEN DIOXIDE; nitrous oxide; oxygen; OZONE; trace elements;
  TRACE GASES; atmospheric pressure; AIR TEMPERATURE; WATER VAPOR;
  upper level winds; solar radiation; reflected infrared;
  solar active regions
```

Fig. 2. Datasets "near" ATMOS dataset.