Earth Sci Inform (2008) 1:29–34
DOI 10.1007/s12145-008-0002-7

METHODOLOGY ARTICLE

A brave new (virtual) world: distributed searches, relevance scoring and facets

Todd King · Tom Narock · Raymond Walker · Jan Merka · Steven Joy

Received: 6 August 2007 / Accepted: 18 January 2008 / Published online: 21 February 2008

Abstract Our ability to deal with complex systems has improved through information system research which includes improved modeling (both data and system), the use of semantics and advances in distributed computing. The past decade has seen an explosion in the amount and variety of geosciences data and the emergence of true open data repositories through which scientists can freely access this data. Those data are found in thousands of repositories located around the world. Virtual observatories have been created to address the challenge of helping scientists search those repositories to find and access the required data. This challenge is been addressed by using technologies such as the Internet (with ample connectivity and bandwidth), the Web, cheap computing power, cheap storage and standards for critical components. Many scientific disciplines are developing virtual observatories. Yet some of the most compelling science questions cross multiple domains. While semantics can provide cross domain reasoning, often the first step in answering a question is determining what resources are available which may be relevant to a topic. The topic can be expressed as simple phrases or word sequences. Using a common relevance scoring method at all locations can enable a federated search across loosely coupled providers. The results of which can be organized into facets to aid the user in selecting the most promising resources with which to pursue the scientific investigation. We describe an approach to developing and deploying relevance scoring methods and faceted results in this brave new (virtual) world. We have found that a scoring method which considers both the presence of terms and the proximity of these terms relative to the order of the terms in the query improves the assessment of relevance. We call this Term Presence-Proximity (TPP) scoring and describe a method for calculating a normalized score. TPP scoring compares favorably with other scoring approaches.

Keywords Relevance scoring facets virtual observatory search

Editorial Responsibility: P. Fox

T. King (✉) · R. Walker · S. Joy
Institute of Geophysics and Planetary Physics,
University of California,
Los Angeles, CA, USA
e-mail: tking@igpp.ucla.edu

T. Narock · J. Merka
Goddard Earth Science and Technology Center,
University of Maryland Baltimore County,
Baltimore, MD, USA

T. Narock · J. Merka
NASA/Goddard Space Flight Center,
Heliospheric Physics Laboratory,
Greenbelt, MD, USA

Introduction

The development of the Internet was based on four basic principles described by Kahn (1972) which can be paraphrased:

1. Each distinct network should be autonomous.
2. Communications should be on a best efforts basis and retried if unsuccessful.
3. Connections between networks must be transparent.
4. There should be no global control at the operations level.

These same principles apply to the web if "network" is replaced with "web site". The web has made a wealth of information available over the Internet. The same is

currently occurring with scientific information. The past decade has seen an explosion in the amount and variety of geosciences data and the emergence of true open data repositories through which scientists can freely access this data. Those data are found in thousands of repositories located around the world. Applying the Kahn Principles to sharing of scientific data leads to what is called a "virtual observatory". A Virtual Observatory is an on-line environment that provides uniform access to data and services for a community which can defined by scientific interest, resources, and geo-politics. Currently, there are virtual observatories in astronomy (Hanisch and Quinn 2003), planetary exploration (Hughes and Yi 1993) and space physics (Harvey et al. 2004). Each of these domains has developed information models, taxonomies and in some cases formal semantics to describe available resources. These are expressed as implementation specific data models which are exposed through services or simple file sharing to enable searching and harvesting. Within each domain resources are described with the data model to allow a scientist to discover, access, analyze, and combine data from multiple sources in uniform, user-friendly ways. However, the discovery process which is the first step in acquiring resources to answer pertinent science questions requires some knowledge of the domain's data model. For some domains such as planetary exploration and space physics there are discipline specific sub-domains so the expertise of a user does not have encompass the entire domain. In general, virtual observatories exist world-wide, operate as peers, and are administered independently.

Often the answers to the science questions being posed require resources from multiple disciplines. This places a burden on the user to become familiar with multiple data models. One approach to minimize this burden is the use of semantics to translate requests posed in the "language" of one domain into that of another domain. This approach requires that translation technologies be deployed but they are not yet widely adopted. An alternative approach is to adopt a federated search architecture where a search is distributed among peer systems which use common methods and data models for comparing short phrases or groups of words to the available resources. The results of these comparisons can then be blended together at the point of request to give the user a coherent picture of the relevance of available resources. One benefit of a federated search is the ability for parallel, distributed execution. Federated searchers can be deployed with minimal intrusion, into the existing data environment where data are distributed and under the control of many people and institutions.

In this paper we will discuss an approach for calculating the relevance of a resource to a set of search terms and ways to organize the results based on facets or common attributes to further aid the user in locating the most relevant resources. Since this approach is domain neutral will we also discuss how the approach can be utilized in federated search architectures analogous to the virtual observatory environment.

Methods for determining relevance

The relevance of an item can be determined by inspecting its attributes and comparing them to the desired attributes. For instance a scientist studying aurorae wants to know if there were changes in the magnetic field near local midnight at the time of auroral observations. In this case relevant data would be magnetic field measurements from stations near midnight at certain magnetic latitudes at the time of the auroral observations. A user interface will allow this query to be expressed in a natural way. For instance a natural language query might be "simultaneous auroral images and ground magnetic observations". The query would then be parsed and compared to the attributes of all known items. Each item may have many attributes, some of which are contextual. For example, attributes might include the time of the observation, the type of observations, and their location. When an item is described the attributes are represented by metadata. We call the item that has been described by metadata a "resource".

In the space sciences there are three domain specific metadata standards for describing resources. In astronomy there is the IVOA (International Virtual Observatory Alliance) data model (Hanisch 2007); in planetary science there is the PDS (Planetary Data System) data model (Hughes and Yi 1993) and in space physics there is the SPASE (Space Physics Archive Search and Extract) data model (Harvey et al. 2008). For a relevance determination method to be effective it must produce reliable results independent of the data model. The method described later meets this requirement. Before we describe this method let us look at some other methods and explore why they are ineffective.

One method of relevance determination called Term Frequency (Salton and Buckley 1988) is based on the premise that a resource is more relevant if a desired term appears often in the resource. Term Frequency is typically normalized to prevent bias based on longer documents. Term frequency for term i (tf_i) is calculated as:

$$\mathrm{tf}_i = \frac{n_i}{\sum_k n_k} \tag{1}$$

where n_i is the number of occurrences of the considered term, and the denominator is the number of occurrences of all terms in the resource. One limitation of the Term

Frequency method is that a "perfect" match ($\Sigma tf_i = 1$) occurs only when every search term is found in the document. Also, when searching across multiple documents of differing lengths additional normalization is needed in order to compare matches between resources. Typically the log of the inverse document frequency (idf; the total number of documents (resources) divided by the number of documents which contain the term) is used. If a document does not contain any terms the idf is undefined.

The relevance scoring determined with the Term Frequency-Inverse Document Frequency (tf-idf) method is applicable only to a single collection of documents. It is not possible to compare the score of an individual document in one collection to a document in another collection since the size of the collection influences the relative score.

Deerwester et al. (1990) describe a technique for latent semantic indexing in which a term-document matrix is constructed to analyze the relevance of documents to each other and to create a "semantic" space which a query searches to locate appropriate clusters of documents. The algorithm for construction the semantic space involves singular value decomposition (SVD) which is similar to the calculation of eigenvectors and eigenvalues for the term-document matrix. As with the Term-Frequency methods the latent semantic indexing technique applies to a known collection of documents.

Another relevance scoring method is one used by Google called PageRank™ (Page 2006). The premise for PageRank is that a resource with more links to and from itself is more relevant. This method has proven to be quite effective for the web where links are a "vote" of relevance. The PageRank is applied to a collection of resources to determine the relative importance of each resource. The collection is obtained through text based searches where formatting of the text influences the scoring (Brin and Page 1998). For example font size and style adjust the scoring. Additional attributes which influence scoring are where a term appears in the document for example, in the title, anchor, URL, the body, etc. When a collection is obtained the PageRank is determined so that the most relevant documents appear first in the list.

The success of Google demonstrates the effectiveness of the PageRank approach in a web environment, but science data is different. Most resources are single observations that are not linked to other resources so the PageRank for all resources is the same. Resources also do not have inherent attributes that are readily indexed and so metadata must be created to make the resources tangible. The metadata are highly structured and typically state an attribute only once so techniques such as Term Frequency degenerate into $1/n$ scores. Also, the structure or tagging inherent in the metadata can be used to give higher importance to some words based on context and can increase the accuracy of this approach. Web-based searches, such as Google, do this to different degrees, but have an "HTML" semantic environment.

The most common approaches to word based searches are not very useful for space science data. For science data the approach should (1) be effective with concise descriptions (common in metadata), (2) have universal scoring (collection independent), (3) be structurally independent (the method of expression doesn't matter) and (4) be semantically adaptive (phrasing of information influences the search). In the next section we present a method of relevance scoring which meets these objectives called "Term Presence-Proximity" (TPP).

The Term Presence-Proximity algorithm

The Term Presence-Proximity score for any resource is determined by first generating a word index for the resource. A word index can be formed by:

1. Scanning the resource description as a stream
2. Converting appropriate tags to words
3. Parsing tag content into words
4. Adding words to the ordered list as they appear and
5. Only adding words the first time they appear.

The resulting word list contains a list of unique words in the order of first appearance.

The presence part of the Term Presence-Proximity score is calculated by comparing each search term to the word list and determining how many of the search terms appear in the list. The value of the presence score is:

$$p_r = \frac{m}{n} \tag{2}$$

where m are the number of matching terms; n is the total number of search terms. The value of the presence score ranges from 0 to 1. For example, if the search terms entered by the user consist of 4 terms and only 2 of the terms occur in a resource then $m = 2$ and $n = 4$ and the presence score would be $p_r = 2/4 = 0.5$. If all terms are found the presence score is unity.

The proximity portion of the Term Presence-Proximity score is calculated by measuring the distance between successive search terms found in the word list. The value is calculated as:

$$P_x = \frac{m}{\sum_{i=1}^{m} |l_{i-1} - l_i|} \tag{3}$$

Where l is the location (index) of a search term in the resource word list and m is the total number of terms found in the resource word list. The first word found is considered

to have a distance of 1($|l_0 - l_1| \equiv 1$). If the search terms are adjacent in the order given then the proximity score will be unity. The more separation between the search terms in the word list the lower the proximity score.

The two parts are summed to determine the Term Presence-Proximity score. If both parts are given equal importance then the relevance score is:

$$\text{relevance} = \frac{m}{n} + \frac{m}{\sum_{i=1}^{m} |l_{i-1} - l_i|} \tag{4}$$

A perfect "match" will have a relevance score of 2. The terms may be weighted differently, but current tests have not provided a compelling reason to do so.

The accuracy of the Term Presence-Proximity score can be improved by making common adjustments to the contents of the word list and applying similar adjustments to search terms. Such adjustments include an exclusion of: (1) Superfluous words (words which are irrelevant for the domain), (2) Articles (a, an, the), (3) Prepositions (from, of, to) and (4) Deixis (this, my, your). In addition specific content such as unique system defined identifiers (resource identifiers) and other information clouds (groupings of information) may be excluded since the information content may be negligible.

Additional terms can be added to the relevance score formula to include external metrics into the final score. For example, the number of times a resource was accessed in a given period time is an indicator of popularity. For science resources popularity increases after results obtained by using a data resource are presented. For our application we have not included other terms such as a popularity score since it may obscure resources for new science questions.

Federated searches using Term Presence-Proximity scoring

Since each resource is evaluated independently of all other resources Term Presence-Proximity scoring can be performed in parallel on disparate collections and then combined into a single set of search results. In a federated environment where each operating unit uses a common scoring method it is possible to implement a distributed or federated search. Figure 1 illustrates the functional components of a federated search system. At each location the Term Presence-Proximity scoring is performed on the local resources. In the illustration only the top 5 most relevant resources are returned. At various nodes the results from other locations are blended together with local results and the top 5 are then selected. Since each score is determined with a common method the final result is a simple merging of the results from all sources, followed by a sorting on the score and selection of the most relevant resources from the new set.

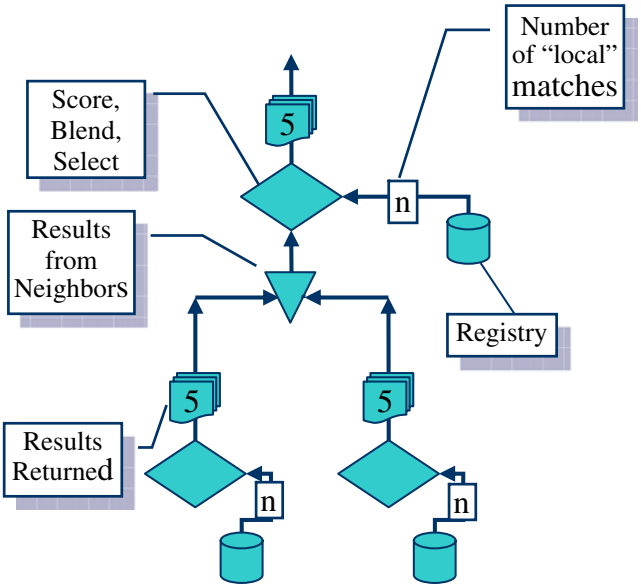

Fig. 1 Illustration of a federated search. Barrels indicate information stores (registries), *arrows* the flow direction of information, *boxes* on each arrow indicate the number of items returned, *inverted triangles* indicate points where information is merged from multiple sources and *diamonds* indicate decision making (such as which results to pass on)

Using facets to organize search results

The display of search results can be more effective if common attributes are used to organize the results. Some examples of common attributes are resource type, classification or common characteristic. A facet can be domain or user specific. For example, one facet could be the provider of the resource or type of resource such as service, data, or document. Term Presence-Proximity scoring can be performed for each facet so the most relevant resource in each facet can be displayed for the user.

Results of tests using the TPP approach

Multi-faceted federated searches are used in a search service provided by the Virtual Magnetospheric Observatory (VMO; http://vmo.igpp.ucla.edu). The facets in this system are the resource types as defined in the Space Physics Archive Search and Extract (SPASE) data model (http://www.spase-group.org/). The VMO also supports a structured search capability (http://vmo.nasa.gov). Tests of multi-faceted Term Presence-Proximity scored using the VMO show that it has promise as a search tool for structured information. For the tests we used the real world resource descriptions in the existing VMO. At the time of the tests there were 63,226 inventoried resources in the system consisting of 1,706 primary resources and 61,520 file level resources associated with a primary resource. These were harvested from 25,367 resource descriptions

(structured metadata stored in XML). Word based searches are performed only on the primary resources because in SPASE a file level resource is an augmentation of a primary resource. The query used to illustrate the effectiveness of the Term Presence-Proximity scoring is representative of the type of query a researcher in magnetospheric physics may ask. The specific query was:

calibrated plasma data in the magnetotail

The results from various queries are shown in Table 1.

Three methods were used to evaluate the response. The first method was no scoring which simply returned those resources which contained any of the terms in the query. The second method used presence scoring only (see Eq. 2). The third set of results were found by using the full Term Presence-Proximity scoring approach (Eq. 4). Each method returned the same number of results (841) This was expected since any resource which contains any of the terms is considered potentially useful. The order of the resources is different with each method. The top ten resources returned are shown in Table 1. For the "no scoring" method resources describing attributes of propagation calculations in the solar wind filled the top six positions. These are not relevant to this query. While this information would be of interest to someone using data to study the solar wind, it is not useful for magnetospheric studies. The seventh and eigth items are relevant since the data was acquired in the magnetosphere, but the ninth and tenth are for groundstations which may be peripherally relevant. In short the results are a jumble and would require more exploration to determine actual relevance. With presence scoring only the relevance of the top ten resources is improved greatly. The results begin to depart after the sixth resource. With the Term Presence-Proximity (TPP) method relevance of the top ten resources is further improved. The top six resources are the same as the with presence only scoring because all terms are present and have relatively the same proximity. For resources seven, eigth and nine three of the terms are present and the proximity of the terms draws the "Geotail Low Energy Particle Data" and "Wide-range 3D Ion Spectrometer" resources into the list. These resource are much more relevant to the query then the corresponding entries in the presence only list.

Discussion

Term Presence-Proximity scoring does not depend on a specific semantic model. The word list normalizes any semantic model into a generic first occurrence word list. It shares some features with the latent semantic discover approach of Deerwester et al. (1990) where a term-document matrix is constructed for a collection of documents. The presence portion of the Term Presence-Proximity method is equivalent to a $1 \times N$ term-document matrix when N is the number of terms. The addition of the proximity term enhances the latent semantic discovery

Table 1 Results for different methods

Query: calibrated plasma data in the magnetosphere

Method	Number	Score	Resource
None	841	0	ISEE-1 Propagation details
		0	IMP8 Propagation details
		0	ACE Propagation details
		0	Wind Propagation details
			Geotail Propagation details
			ISEE-3 Propagation details
			Polar Magnetic Field Experiment (MFE) Data
			Comprehensive Plasma Instrumentation data
			Panagyurishte Fluxgate Magnetometer Data
			Zaymishche Fluxgate Magnetometer Data
Presence	841	1,000	Comprehensive Plasma Instrumentation data
		1,000	ISEE-3 Fast Plasma Experiment
		1,000	Solar Wind Plasma Faraday Cup data
		1,000	ISEE-2 Fast Plasma Experiment
		1,000	Wind 3DP
		1,000	ISEE-1 Fast Plasma Experiment
		750	ISEE-3 Propagation details
		750	ACE SWEPAM
		750	ISEE-3 Tri-axial fluxgate magnetometer
		750	Fluxgate DC Magnetometer FM-3I data
TPP	841	785	Comprehensive Plasma Instrumentation
		785	ISEE-3 Fast Plasma Experiment
		785	Solar Wind Plasma Faraday Cup data
		785	ISEE-2 Fast Plasma Experiment
		785	Wind 3DP
		785	ISEE-1 Fast Plasma Experiment
		675	ACE SWEPAM
		675	Geotail Low Energy Particle experiment data
		675	Wide-range 3D Ion Spectrometer (CORALL)
		625	Fluxgate DC Magnetometer FM-3I data

A score of 1,000 is considered a "perfect" match. A score of 0 indicates no score was calculated. Each method returned the same number of results. The top ten resources returned are shown in the table. From a researcher's perspective the top results returned with the Term Prsence-Proximity (TPP) method are more relevant since they are resources which relate more directly to the useful plasma observations.

method by considering the order of terms in the scoring. The Term Presence-Proximity method provides additional benefits since it combines the initial selection criteria of term presence with a term-to-term proximity score which acts like pseudo-semantic scoring. This results in a simple, efficient and reasonably accurate score which can be used for resource selection in a federated search environment. Current uses of the Term Presence-Proximity have shown that the method is fast and easily focused by the user since it places terms in a natural order and since using more search terms increases the relevance of the results.

Conclusions

Adopting a common relevance scoring algorithm makes possible efficient federated searches. The generation of the word lists for each resource is independent of the underlying data model or metadata environment. This enables meaningful federated searches and improves the accuracy of results. Adopting common facets provides users with greater selection control. The facets can be based on the data model used in a domain so that the division of results reflects a "natural" organization within the domain. Federated searches can be very effective in a virtual observatory environment where data sharing is highly desired, but where institutional and political boundaries require local control over the resources. While different scoring methods may be used in federated searches we have found that a scoring method which considers both the presence of terms and the proximity of these terms relative to the order of the terms in the query improves the assessment of relevance for each resource. The Term Presence-Proximity method is a simple and efficient approach which may be applicable to other domains.

Acknowledgements This work was supported by the National Aeronautics and Space Administration under Grants No. NNX07AC95G and NNX07AC93G and issued through the Virtual Observatories for Solar and Space Physics Data (S3CVO). The UCLA/IGPP publication number is 6360.

References

Brin S, Page L (1998) The anatomy of a large-scale hypertextual web search engine. Stanford InfoLab Publication Server

Deerwester S, Dumais ST, Furnas GW, Landauer TK, Harshman R (1990) Indexing by latent semantic analysis. J Am Soc Inf Sci 416:391–407. http://citeseer.ist.psu.edu/deerwester90indexing.html

Hanisch R (2007) Resource metadata for the virtual observatory version 1.12., IVOA Recommendation 2007 March 2. http://www.ivoa.net/Documents/REC/ResMetadata/RM-20070302.pdf

Hanisch R, Quinn P (2003) The International Virtual Observatory. http://www.ivoa.net/pub/info/TheIVOA.pdf

Harvey CC, Thieman JR, King T, Roberts DA (2004) SPASE—Space Physics Archive Search and Extract, PV-2004 ensuring the long term preservation and adding value to scientific and technical data, Frascati, Italy

Harvey CC, Gangloff M, King T, Perry CH, Roberts DA, R Thieman J (2008) Recent developments towards a Solar System Virtual Observatory. Earth Science Informatics (in press)

Hughes JS, Yi YP (1993) The Planetary Data System data model. Twelfth IEEE Symposium on Mass Storage Systems, IEEE, Monterey, CA, pp 183–189

Kahn R (1972) Communications principles for operating systems. In: Internal BBN memorandum

Page L (2006) Method for node ranking in a linked database, Patent number: 7,058,628, The Board of Trustees of the Leland Stanford Junior University, Stanford, CA

Salton G, Buckley C (1988) Term-weighting approaches in automatic text retrieval. Inf Process Manag 245:513–523