Meta-Information for Knowledge Navigation and Retrieval: What’s In There

David Steier, Scott B. Huffman, and Walter C. Hamscher
Price Waterhouse Technology Centre
68 Willow Road
Menlo Park, CA 94025
steier@tc.pw.com

Introduction

Knowledge navigation in the business world is often a difficult and expensive process. Commercial databases are notoriously incomplete and inaccurate, access charges can run into hundreds of dollars per hour, and the client always wants the answer yesterday. In coping with such problems, professionals who search for information on a daily basis rely on meta-information that goes beyond standard descriptions of the content of data sources. For example, in working with information specialists at Price Waterhouse LLP (PW), a large professional services provider, at one point we were advised, “If you’re looking for quick information about foreign companies that’s not too out-of-date, use database so-and-so because the vendor is offering free connect time this month.” Such a heuristic does involve the source’s semantic content, of course, but also refers to resources (money) available for the retrieval, and quality (timeliness) of the retrieved data. If the tools we build are to make knowledge navigation in the business world any easier, they will have to exploit, or at least represent, this type of meta-information in choosing information search and retrieval strategies.

A number of researchers have noted the importance of meta-information. KQML [Finin et al., 1994] provides an interprocess communication language for information agents, but makes few commitments about the semantics of the meta-information to be communicated. MCC’s Carnot project [Collet et al., 1991] introduced some of the necessary computational infrastructure with concept matching and resource models. The SIMS project [Arens et al., 1993] exploits some semantic and run-time meta-information to produce efficient query plans.

Yet there has been little reported in the literature on exactly what meta-information is important for describing real-world task contexts. We present our preliminary findings on this subject in this paper. Our goal is to create a “content theory” on which to base languages and support tools for the delivery of knowledge services, in contrast to focusing more narrowly on answering queries in isolation. We define a knowledge service as the provision of content in some form and at some level of quality, subject to constraints on the resources needed to perform the service. This paper begins with some conceptual preliminaries, and then discuss these four elements in more detail. As illustrative examples of the use of meta-information, the paper describes two knowledge services: monitoring management changes, and search for potential conflicts of interest in prospective litigation support clients.

Conceptual Foundations

The field of heterogeneous database integration is a natural starting point for understanding the use of meta-information, but the topic has received only limited attention there, especially in commercial database tools. Some research prototypes represent meta-information for field- and schema-level integration, where for example monetary units are represented in different currencies [Sciore, et al, 1994]. Although this work is certainly relevant, a higher level of abstraction than is available in traditional database terminology may be needed to describe meta-information about content, form, quality and resources for knowledge services. We therefore follow SIMS [Arens et al., 1993] and Carnot [Collet et al., 1991] in shifting the level of abstraction upwards from relations among database attributes to relations among concepts. Here the term concept corresponds to the informal standard sense as referring to an abstract entity. Concepts can be defined recursively in terms of other concepts, and all concepts refer to some central domain ontology that is assumed common and is used to arbitrate the semantics of a particular reference. One can view even unstructured information sources such as newswires as providing instances of a concept such as [Event] 1. For each concept, there is mapping knowledge that can be used to translate a concept into some representation in a particular source, and vice versa. This abstraction level is essentially the “knowledge level” [Newell, 1981] specialized to knowledge navigation.

One could imagine an extended SQL style of notation for specifying a knowledge service’s concepts

---

1In this paper we will enclose concept names in brackets.
and associated meta-information. Parameters in SELECT clauses would refer to concepts being retrieved, AS clauses would specify the delivery format, FROM clauses would describe sources and so on. A knowledge service specification would thus take the form:

```
SELECT [concept] FROM [source list]
AS [form constraints]
WHERE [content constraints]
SATISFYING [quality constraints]
USING [resource constraints]
```

Next, we turn to two case studies of knowledge services within PW, with an eye toward identifying the space of values for each of these meta-information clauses.

Case Studies

The following two examples will be used to illustrate the meta-information categories we have identified. The first is taken from a knowledge service we support with tools now in routine use at PW, while the second is currently performed with little automated support, and might be a target domain in the future.

Case Study #1: Management changes

Our first case study involves the filtering of newswires for news of management changes. The main input data source is a newswire service called NewsEdge (TM), which on a typical day will deliver 5000 articles of business news. Articles reporting management changes contain sentences like:

"Lawrence P. Murphy has been named executive vice-president and chief strategic officer of The Walt Disney Company (NYSE: DIS), Michael D. Eisner, chairman and CEO announced today."

The simplest strategy for filtering would be to look for keywords that are likely to indicate a management change; unfortunately, in this case such searches return on the order of 1000 hits per day. Of these, typically only about 100 indicate genuine management changes, and only a fraction of the full text of those 100 articles give the needed information. Thus, this task is a prime candidate for intelligent support. The information of interest in the task is the name of the person, the name of the company, and the person's new job title. For the example sentence above, an automated system must know enough to report Murphy's name and title, rather than Eisner's. To perform this task we built an information extraction system called ODIE [Huffman, 1995], which looks for shallow linguistic patterns using local syntactic relationships.

The second example involves searching PW records for potential conflicts of interest before we take on a new client. As part of its expansion beyond its traditional role as one of the “Big Six” accounting firms, PW performs substantial amounts of litigation support. In a legal dispute, PW cannot work for both sides, so the firm must check if it is already performing some service for an entity related to another party to the litigation, before accepting a new litigation support client. This check for conflicts, often performed under severe time pressure, must cover a diverse range of relationships (both direct and indirect) that could establish a conflict - ownership, directorships, competitive position, etc. A variety of sources that record such relationships (some on-line, some not) are consulted to assure that no conflict exists before a client is accepted. If a conflict is missed, the potential costs to the firm are huge. However, because only one conflict is enough to rule out a prospective client, the utility of finding additional conflicts after the first one has been found is very small. The general strategy used for this task by the information specialists in our London office is to first generate a list of entities (officers, subsidiaries, etc.) related to the prospective client, and then to check for conflicts with each of those entities, and finally send a mail message to the person requesting the search with the results, listing each PW involvement found.

The four categories of meta-information

In both case studies, providing a knowledge service can be described as a flow of concepts. Concepts originate from one or more sources, are processed or filtered in some way, and sent to a sink for storage or delivery to the service requestor. Each of the types of meta-information we have mentioned: content, form, quality, and resources, describe an aspect of the flow of concepts in a service. Drawing from the case studies, we describe each of these four types in more detail below.

Content

Content meta-information describes a service with respect to the semantics of the concepts delivered by the service. The service that monitors management changes accesses a source that provides instances of the [Event] concept and sends out a stream of instances of the concept [Management Change]. Similarly the second service example uses a variety of concepts as input and delivers [Litigation Support Conflicts].

The meta-information about concepts delivered by a service can be very specific. For example, we found that the particular newswire we used as a source for
management change monitoring does not contain articles reporting events in companies outside the U.S. That fact is useful meta-information for our international affiliates who might consider using the service. We've found no substitute for a well developed conceptual ontology, one that represents every useful distinction in a domain. The ontology must be shared by those who specify requests for services and the creators of meta-information for candidate services to fulfill that request.

Form
Form meta-information specifies how the sink in the knowledge service delivers or stores concepts for the service requestor. Examples are:

- **Medium**: In the management changes service, the sink is a Notes database, while in the conflict search service, the output is electronic mail. Often in applications such as the conflict search, some of the input sources are in books or reports (yes, on paper), and outputs are letters or FAXes.

- **Format**: If the delivery medium is a Notes database, the format meta-information includes the forms and fields on each form in the database. The “Management Changes” database contains a form with fields for the date of the change, the name of the company, the new position, the person appointed to the position, the source text, and pointers to documents in other relevant databases.

- **Indexing/Ease of access**: Meta-information about the format of a knowledge service also includes information about how easy it is to access various instances within the sink. That is, what kinds of indices are available? Are instances pre-classified into useful categories within the sink? Are text search tools or other kinds of query engines available within the type of medium? The management changes database, for instance, classifies its instances by the company involved, into categories of interest within PW (geographic region, company size, industry, etc.); in addition, Notes provides full-text searching over the database.

- **Volume**: The frequency of delivery of the results of a service can be an important determinant of how usable the service is (for example, mailing lists). The users of the conflict search service want essentially one message per search request. The management changes database typically grows by 75 to 100 entries per business day, necessitating the types of indexing described above. The indexing provides an additional filter to users, so that the typical user (who is interested in a particular region and industry) only examines a few entries of interest per week.

- **Access and redistribution rights**: Commercial information providers often limit access to subscribers, and prohibit redistribution of information to other potential clients of their services. There are often geographic limitations on redistribution as well, specified in contracts with particular data sources.

Quality
Meta-information about quality describes how reliably a concept can be produced by the service. Instances include:

- **Recall and precision**: The information retrieval community has long measured completeness and accuracy of retrievals in terms of percentages, and these terms can be useful in describing knowledge services. For example, we are achieving 80–85% recall in the management changes monitoring service, with about 90% precision.

- **Timeliness**: The expected lag between an event’s occurrence and the time the user is notified of it must be small enough for the information to be of value to the user. The management changes timeliness is measured on the order of days. Complementing the need to include the most recent information is historical coverage: the conflict search must go back several years to be complete.

Resources
There is a diverse set of types of meta-information about resources needed to use a service. Resource constraints can be expressed as absolute limits (don’t spend more than $500 or three hours getting this information) or as a budget over time (don’t spend more than $100 a week). There are also a variety of methods of pricing information: charges may be assessed per user, per location, per search, per records examined, or per pages printed.

In the conflict search example, there are a number of sources (we’ve counted 10 so far) that might be consulted to find conflicts. Because one positive result is sufficient to terminate a search, it pays to try the cheapest source first. If no conflicts are found, specifying the stopping criterion for the conflict search problem can be difficult. In most cases, stopping is based on time or cost limits, but in some cases information specialists use their knowledge of the sources to identify the point when the search is reasonably “exhausted”.

Resource meta-information is not limited to monetary cost. Other examples include:

- **Skills**: Meeting the specific needs of a service request may require developers and end-users with certain skills. For example, users of our management changes database must have basic Notes familiarity. Information professionals who use on-line services are assumed to know the common search language used by Dialog and other providers.

- **Time**: Turnaround time from request to delivery is the most common type of time resource meta-information, but there are some unusual twists. In
the conflict search example, sometimes people in overseas offices need to be contacted, and because of time zone differences this will introduce a lag into the process. In a less routine situation, time may have to be allocated for service development or user training, depending on the familiarity with the service.

- Minimum hardware requirements: A service may require minimums of processing power, primary and secondary memory, communication bandwidth, etc.
- Operating system platform: Most business information sources on CD-ROM are targeted for DOS or Microsoft Windows, which may preclude easy access by users of other platforms.
- Application or database software needed: Our management changes service is only available to users of Lotus Notes; similarly, some services provide documents that can only be read by particular word processing software.

Implications

Each of the four categories of meta-information described above can have an impact when navigating among knowledge services. Unfortunately, the sheer diversity of the types of information we have listed does not bode well for fully automated knowledge navigation among commercial providers of knowledge sources. Even if we invented a language capable of expressing this meta-information, there are still major issues with respect to how meta-information is disclosed and acquired. For example, access and redistribution rights, which can be critical issues in selecting a datasource, are normally expressed in complex legal terms in contracts available only in printed form. For commercial on-line vendors, objective standards of data quality and search effectiveness are nearly non-existent, and there may not be any economic incentives for disclosure of such meta-information in a machine-readable form.

Yet, it is evident that human information specialists somehow manage to acquire and use meta-information from all four of the categories. A useful short-term goal for knowledge navigation research might be to design representations and tools that simply record meta-information, and make it explicitly available for future searchers. A simple search of a database containing this information may be sufficient to help identify sources and plan for the construction of new knowledge services.

Acknowledgements

We are indebted to the information specialists of Price Waterhouse for helping us understand what makes knowledge navigation so hard, and especially to the staff of the Information Centers in Menlo Park and in London for help with the management changes and conflict search examples.

References


