Using Hybrid Search and Query for E-discovery Identification

Dave Grosvenor and Andy Seaborne

Hewlett-Packard Laboratories, Bristol
{dave.grosvenor,andy.seaborne}@hp.com

Abstract. We investigated the use of a hybrid search and query for locating enterprise data relevant to a requesting party's legal case (e-discovery identification). We extended the query capabilities of SPARQL with search capabilities to provide integrated access to structured, semi-structured and unstructured data sources. Every data source in the enterprise is potentially within the scope of e-discovery identification. So we use some common enterprise structured data sources that provide product and organizational information to guide the search and restrict it to a manageable scale. We use hybrid search and query to conduct a rich high-level search, which identifies the key people and products to coarsely locate relevant data-sources. Furthermore the product and organizational data sources are also used to increase recall which is a key requirement for e-discovery Identification.

Keywords: SPARQL, e-discovery, identification, hybrid search and query.

1 Introduction

E-discovery is the process of collecting, preparing, reviewing, and producing electronic documents in a variety of criminal and civil actions and proceedings [1]. In this paper we address the problems of scale and recall in the *identification* stage of e-discovery which is responsible for learning a coarse location of data relevant to a legal case. There are two components to our approach to these problems. The first component was to add search directives to SPARQL [2] to give two different information retrieval models in a hybrid search and query. This gives integrated access to both structured and unstructured data sources in the enterprise. However the second component was to exploit some common product and organizational data sources to both guide the searches to cope with scale, and to increase recall.

This paper is organized into five sections. Firstly we extend the introduction with an overview of e-discovery, identification and related work. Secondly we motivate our approach to hybrid search and query. Thirdly we discuss our use of some common data sources to both guide the search, and restrict it to a manageable scale. Fourthly we examine a hypothetical patent violation e-discovery case and give examples on the use of hybrid search and query. Finally we give our conclusions on the investigation.

A. Bernstein et al. (Eds.): ISWC 2009, LNCS 5823, pp. 811–826, 2009.

1.1 E-discovery

E-discovery is a new issue for enterprises created in 2006 by the Federal Rules for Civil Procedure (FRCP) [3] in the US which legally require enterprises:

- To disclose the identities of all individuals likely to have discoverable information relevant to a legal case.
- To either provide a copy, or the location and description of all Electronically Stored Information (ESI) relevant to a legal case.

The courts can potentially impose punitive damages on an enterprise for a failure to comply.

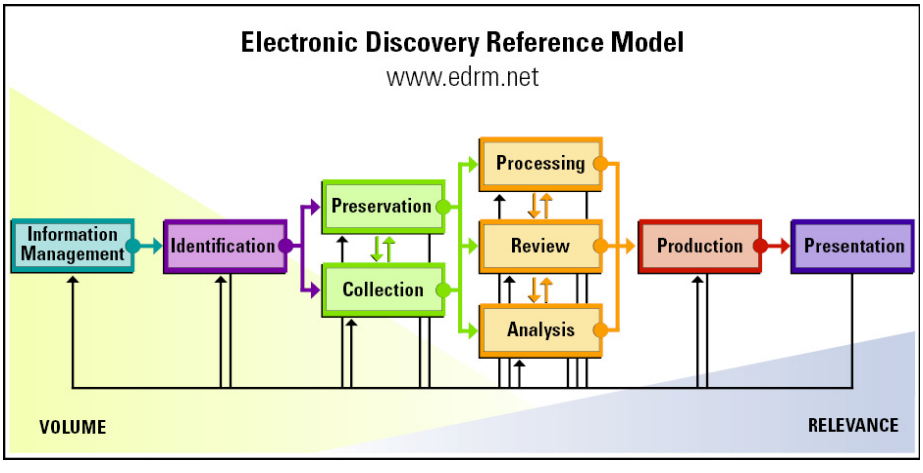

Fig. 1. Reference Model for E-discovery

The ESI includes both structured and unstructured information within the enterprise which specifically includes e-mails, web-pages, word processing files, and databases stored in the memory of computers, magnetic disks (such as computer hard drives and floppy disks), optical disks (such as DVDs and CDs), and flash memory (such as "thumb" or "flash" drives). Currently email is the most important source of discoverable information (80-90% according to a Magistrate Judge Survey [4]). For example email analysis was used extensively in the ENRON case and is a subject of the TREC legal track [5].

E-discovery requests can be initiated by arbitrary legal cases which makes it difficult to prepare for in advance. According to the "Magistrate Judge Survey" [4] only a few types of case are responsible for most of the e-discovery issues. These types of cases were: Individual plaintiffs in Employment cases; General Commercial cases, Patent or Copyright cases, Class Employment actions, and Product Liability cases. The Federal Judicial Centre provides examples of e-discovery requests [6], preservation orders [7], and 'meet and confer' forms [8]. The Sedona conference [9] has been influential in the development of approaches to e-discovery. In this paper we examine a fictitious patent violation case.

E-discovery is a complex guided search task conducted over a long duration (perhaps several months) by an expert team consisting of legal and IT experts acting on behalf of the disclosing party. The *Electronic Discovery Reference Model* (EDRM) [10] has been developed to explain the different stages occurring in the e-discovery process. Figure 1 shows the sequence of stages and indicates that the earlier stages must process a greater volume of data with each unit of data being unlikely to be relevant to a request, whereas the later stages process less data but with each data unit is being more likely to be relevant.

The reference model decomposes the e-discovery process into:

- An *information management* stage responsible for all preparation prior to an e-discovery request, including records management and policies.
- An *identification* stage responsible for providing a broad characterization of the data relevant to the discovery request.
- A *preservation* stage responsible for ensuring the ESI identified in the previous step is protected from inappropriate alteration or destruction.
- A *collection* step responsible acquiring the identified ESI.
- A *processing* stage where the volume of data is further reduced and converted into a suitable format for *Review* and *Analysis*.
- A *review* stage where a disclosing party's legal team sorts out both the responsive documents to produce, and the privileged documents to withhold.
- An *analysis* stage where the collection of materials is evaluated to determine relevant summary information, such as key topics of the case, important people, specific vocabulary and jargon, and important individual documents.
- A *production* stage where the ESI is delivered to the requesting party in both an appropriate form and by an appropriate delivery mechanism.
- A *presentation* stage where the ESI produced is displayed before legal audiences (dispositions, hearings, trials, etc..).

Search technology is used for many stages of the reference model [11]. However e-discovery is not just an enterprise search application for two reasons. Firstly e-discovery is not just concerned with the retrieval of documents. The FRCP requires the identity of every person likely to hold relevant information. Furthermore during the identification stage (which we address) is concerned with finding not only specific documents, but also relevant people, projects, organizations and data repositories. Secondly the emphasis in e-discovery is on returning all potentially responsive data whereas enterprise search returns a selection of the documents most likely to be relevant to a request. Precision and recall are two measures of information retrieval performance [12] that are commonly used.

- Precision is the fraction of retrieved documents which are relevant.
- Recall is the fraction of the relevant documents which have been retrieved.

Enterprise search is primarily concerned with precision and typically focuses on finding the few documents most likely to answer a user's question. Whereas e-discovery is concerned with recall and all responsive ESI must be located.

1.2 Identification

Identification is responsible for the initial preparation that allows later automated searches to be performed. It is usual for identification to have a strong manual element and it can be completely manual. It starts from the subpoena and interviews of potential custodians which are performed by the e-discovery team (referred to as custodian-led identification). Identification begins with the key players because the data of individuals who played a central role in the dispute are likely to contain the majority of information relevant to the dispute. An aim of the interviews is a basic framing of the request such as the relevant time frame, the impacted products and organizations, the key witnesses and custodians of the relevant data sources needed. But they also collect information needed for driving later automated searches, such as keyword lists, special language, jargon, acronyms, and technical terms. It will also include creating a data map showing the type and location of the disclosing company's data sources relevant to the request.

Our aim was to provide automated support for identification. We introduced a hybrid of search and query to support a rich model of the identification search task which not only retrieves documents, but people, organizations and products. However identification poses two problems to the use of retrieval technologies:

- The scale of the data which is potentially retrievable during identification.
- The requirement for high recall.

During identification any person, product, organization and data-source within the organization is potentially within its scope. This scope makes the brute force use of search technology more difficult during identification. An output of the identification stage is a manageable selection of the enterprise data that is potentially relevant to the e-discovery request. Only the data selected in identification is subjected to the more detailed review and analysis during of the later e-discovery stages. The recall obtained during identification provides an upper bound for the overall recall during e-discovery and so high recall during identification is very important. As a result identification casts a broad net in characterizing what constitutes relevant data.

1.3 Related Work

There are many approaches to hybrid search and query to which we provide some brief pointers. Research has aimed at search-like retrieval to access structured data, providing keyword search, imprecise queries, top-k selection and ranking [13]. Similarly other research has aimed at more query-like retrieval from unstructured data, using information extraction to provide: querying of unstructured text [14] and searches returning entities rather than just documents [15]. This research simply underlines the value of both models of information retrieval and our approach has been to make use of both. This pragmatic approach has been followed previously, such as in the WSQ/DSQ work [16] which combines the query facilities of traditional databases with existing search engines on the Web. WSQ/DSQ leverages both results from Web searches to enhance SQL queries over a relational database, and uses information stored in the database to enhance and explain Web searches. Parametric search (such as supported by Lucene [17] and most enterprise search engines)

provides a similar capability to hybrid search and query by both associating metadata fields with each document that is indexed, and allowing queries to retrieve their value and to restrict searches to particular values.

There is closely related work [18][19][20][21] which adds full text search capability to SPARQL. They all are concerned with searching the literal strings in RDF datasets. The Sesame like-operator [18] simply filters the results using regular expression matches on the literals in the result sets. Virtuoso [21] provides a system of rules for selecting which RDF triples are indexed. We use an ARQ mechanism for extending SPARQL that allows us to access arbitrary indexes and are not restricted to RDF datasets. In addition ARQ provides support for both full text search of RDF datasets, and selectively indexing RDF triples using Lucene [17].

There are standard approaches to increasing recall using domain knowledge, such as query expansion [22] and spreading activation [23]. Both are automatic means of obtaining more responses to an original query using domain knowledge. This is important for e-discovery identification. Query expansion operates in the query space and transforms the query using the domain knowledge and co-occurring terms to find related or more general search terms and constraints. Spread activation operates in the result space and uses the initial results as seeds that are used to activate other related concepts during a propagation phase. We use neither technique, but we use the product catalog and the organizational data to identify related people and products which are used both in additional queries, and to generate further results.

2 Our Approach to Hybrid Search and Query

In this section we describe our approach to accessing structured and unstructured data using a hybrid of search and query. We give our motivation for using SPARQL extended with search directives, and explain how the search directives are evaluated within a hybrid query.

2.1 Motivation

To exploit both structured and unstructured data sources for e-discovery requires some form of information integration. The semantic web provides useful technology for such integration. We use RDF as common data model to integrate some diverse enterprise data including organizational and product information. SPARQL is used as the common query language. This approach provides a low cost of entry allowing you to query and navigate RDF instance data without the need for semantic integration. This is important as e-discovery potentially requires ad hoc integration to bring together data sources for the particular legal request that would not be used together during the normal operation of the business. Pragmatically we chose to extend SPARQL with search directives to retrieve unstructured documents because search is the predominant means of retrieving unstructured documents.

2.2 SPARQL

SPARQL [2] is a standard query language, protocol and XML result set format as defined by a W3C working group. It became a W3C recommendation in January

2008. A SPARQL query consists of a graph pattern and a query form (one of `SELECT`, `CONSTRUCT`, `DESCRIBE`, `ASK`). The simplest graph pattern is the basic graph pattern (BGP), a set of triple patterns that must all be matched. This is an extension point of the language and we utilize it to add semantic relationships which are not directly present in the data. In particular, we use other indexing technologies, such as free-text indexes to relate text query strings with document URIs.

2.3 Property Functions

ARQ [24] is a query engine for Jena [25] that implements SPARQL. ARQ provides *property functions* as a way to extending ARQ while remaining within the SPARQL syntax, and new capabilities can be added by applications for local needs without needing to modify the ARQ query engine code base. A property function causes some custom code to be executed instead of the usual matching of a triple pattern against the graph data. ARQ executes any property functions in a way that respects the position of the property function in the containing BGP so which variables are already bound at that point in the query does not change.

2.4 Free Text Searches

The property function mechanism has been used to provide access to different indexing technologies, including Lucene, Autonomy Enterprise Search, Google, and Wikipedia. ARQ itself does not provide the free text indexing but provides the bridge between a SPARQL query and the index. The property function implementing the search directives takes a search string and accepts other parameters for controlling the search and return values that are RDF terms. This is usually the URI of the document. In this case the indexing technologies use the body of some arbitrary document as the indexing text which is not part of the knowledge base itself.

3 Data Sources

We address the problems of scale and recall posed by identification by using some common structured and semi-structured data sources both to guide the searches to restrict the scale, and to increase the recall. This use of particular data sources contrasts with the generic but document centric approach followed by the EDRM reference model which is suitable for arbitrary e-discovery requests on arbitrary data sources.

We use some common structured data sources giving the organizational hierarchy and product catalog:

- The organizational hierarchy provides personal contact information for all people working within HP together with information about the reporting structure and a high-level business area names of the organizational structure.
- The product catalog is used by different content management systems within HP to provide different kinds of product related information ranging from product specification to collections of unstructured documents about products intended for use by sales or marketing.

- The product catalog and organizational hierarchy have common fields allowing connections between people and products to be made. For example, the products business area can be used to identify the high-level organization responsible for a product line.

The semi-structured data sources are important because they provide connections between the structured and the unstructured data sources. They connect structured entities to unstructured text which can be used to characterize topics for the search of other unstructured data sources. For example we can retrieve the technical reports written by a particular author to provide document text which can be used to characterize topics. Semi-structured data sources also connect unstructured text to structured entities which can be joined with other structured data sources. So unstructured text in the semi-structured data source can be searched and the entities of the responsive documents retrieved. For example we can find documents responsive to a topic and return the authors of these documents.

- There are many different content management systems within HP which are used to generate the external HP web site, organize unstructured sales brochures and support information. They associate the product catalog with many different forms of structured and unstructured data. They are an important data source for e-discovery.
- There are several repositories of technical reports for which author, creation date, abstract structured fields are maintained. Some of these technical reports are grouped by business area and maintain a record of a report's reviewers.
- The email repositories are very important semi-structured data sources for most e-discovery cases. But for an organization the size of HP they are costly to search without narrowing the search down to particular people and time intervals. They associate people and time intervals to unstructured text and titles which can be searched.
- Patent repositories are semi-structured data sources with structured fields linking people, publications and other patents.

4 An E-discovery Example

In a fictitious case HP is alleged to have infringed a patent on the use of *impressive print technology* assigned to *Another Photo Print Company*. HP is required to disclose information relevant to: the development and use of this technology in its products; estimating the likely profit associated with the use of this technology; showing how sales and marketing made use of the technology. HP is the disclosing party in this e-discovery case, and Another Photo Print Company is the requesting party who initiated the subpoena. HP has their own research and development program for print technology and so whilst complying with the e-discovery request HP is also keen to establish any prior art on the development of such a technology.

The subpoena triggers a duty to disclose and preserve all information relevant to the patent violation case. A team is assembled which is responsible for satisfying the legal request and applying legal holds to preserve relevant data. An identification process is initiated to identify the key witnesses, custodians of data sources, and

finding the location of data relating to individuals and organizations. The initial problem is to get a better characterization of the topic, the people, organizations and products relevant to the case.

The patent alleged to have been infringed will be cited in the subpoena together with some related patents and cited publications. These cited documents can be used to provide an initial characterization of the topics relevant to the case (e.g. as sets of keywords and phrases). This obtains an initial characterization of the technical areas related to the impressive print technology. Similarly the authors of the cited documents are identified and provide some an initial set of people to seed our searches.

Our approach is to use the structured and semi-structured data sources to expand and corroborate the people, products and topics related to the case. We show examples of simple tactics for deriving a set of entities from other entities. These simple tactics can be composed with others to obtain more complex tactics. There is a need for an e-discovery environment to: manage the entities retrieved by such tactics; support the composition of complex tactics; and record the evidence of how they relate to the legal case.

4.1 Finding Relevant Products

The alleged patent infringement is concerned with printing, but the subpoena did not identify the products that may have used the impressive print technology. The e-discovery process must identify these products because the potential value attributed to the impressive print technology needs to be assessed. The subpoena did cite some patents and publications which can be used to characterize some relevant topics. So we use a tactic to find products using these initial topics. The tactic searches for web pages on the HP site for product names and numbers co-occurring with terms related to one of these topics.

This tactic uses the Google search engine to perform a search of the HP public web sites dedicated to product sales and support. Alternatively we could perform a search of the internal content management system which generates the content for this web site. Throughout this paper we will use both parameterized queries to represent such tactics, and the convention that the variable for the parameter is prefixed with a dollar sign and those prefixed with a question mark are bound by the query evaluation.

```
select ?product ?doc{
     ?product product_catalog:product_name ?name.
     ?product product_catalog:product_number ?number .
     ?query ext:printf(
          "site:http://h10010.www1.hp.com %s %s %s"
          $topic ?name ?number).
     ?doc ext:GoogleSearch(?query $n)
   }
```

The example uses a property function performing `printf` function to assemble a query string composed from the topic string parameter together with the product names and product numbers which are retrieved from the product catalog during evaluation of the query. The `GoogleSearch` property function calling the Google

search engine takes a query string for the search and a parameter controlling the maximum number of results to be returned. The Google site query requires all terms to be present for a web page to be returned. So not all searches will return any results and so the search will select products for which there are documents responsive to the topic string. This tactic provides a means of obtaining products related to the topic.

Documents matching the `GoogleSearch` are returned in the (ranked) order that Google returns them. But this is not used to rank the products returned as this would require merging the relevance of scores from distinct searches. Although in this example it would be plausible to do so because each search shares the same terms contained in the topic. It also has some different terms because of the product name and product number. For example, we might want to return the products mentioned in documents that are most responsive to the topic search. But we might also take into account the number and relevance of the documents which are matched.

The ranking problem is better illustrated with a similar tactic which retrieves the products mentioned in documents created by a given person in the technical reports database.

```
select ?product ?doc ?score{
     ?product product_catalog:product_name ?name.
     ?product product_catalog:product_number ?number .
     ?query ext:printf(
          "%s %s"
          ?name ?number).
     (?doc ?score)
          ext:techreportsTextSearch(?query $n $rel).
     ?doc techreports:author $author .
   }
```

This tactic takes three arguments: the author, the maximum number of results, and the minimum relevance of a document. It uses the product catalog to generate query strings using just the product name and number. A search directive uses the generated query string to search the technical reports database. The responsive documents retrieved by the search directive are only returned by the tactic when they were created by the given author. The documents retrieved by the search directive are returned in ranked order for each query string. But for different products the query string is different and so it is not meaningful to use the relevance score of a document to rank the products. We do not address the ranking problem in this tactic, but return additional information with the product result that can be used to discriminate the returned products. We return the document URI and its relevance score. This would then allow someone to discriminate the products returned by arbitrary means, such as the number of responsive documents and their relevance information, or use other attributes of the documents such as date and topic, or whether other products were mentioned in the same documents. We did not address the ranking issue because discrimination of the results seemed sufficient for identification where we were more concerned with recall than obtaining the top-k most relevant results. However the example does illustrate the general problem of ranking the results of such hybrid search and queries.

4.2 Expanding the Set of Products

Once we have identified an initial set of products we can find related products using both the product catalog and the organizational hierarchy.

The product catalog indicates products that are related through being part of series of products addressing a market. For example there will be a printer targeting the consumer market and others targeting the small business office. Over time there will be a series of products addressing this market. Even within these markets products will address different price points and have different specifications. So we can expand from an initial seed product to retrieve all the other printers that are part of the same series or which address the same market.

The product catalog also gives information about how the product is made and where it is supported. For example, every product has a product line which is the responsibility of some organization. This allows us to group products using the organizational structure as well as the market structure. So we can expand from an initial product seed to retrieve all the products that are produced by the same product line and organization. Furthermore we can use the organizational hierarchy for further expansion. For example in the tactic below a product line is followed to its organization which in turn is followed to the business unit, then the direction along the hierarchy is reversed to find all the product lines and products produced by this business unit.

```
select ?product {
  $seed_product product_catalog:product_line ?seed_pl.
  ?pl_org hp_ba_hierarchy:product_line ?seed_pl.
  ?pl_org hp_ba_hierarchy:business_unit ?unit.
  ?org hp_ba_hierarchy:partof ?unit.
  ?org hp_ba_hierarchy:product_line ?pl.
  ?product product_catalog:product_line ?pl }
```

4.3 Finding Relevant People

The subpoena provides an initial characterization of the topics related to the impressive print technology. We now examine a tactic for using a topic to obtain a set of relevant people using a semi-structured data source. We use the HP Labs technical reports repository which we have indexed using both the Autonomy Enterprise search engine and Lucene. Semi-structured data sources, such as the technical reports repository and the content management systems, are very important because they allow the structured and unstructured data sources to be used together.

```
select ?person ?doc ?score {
  (?doc ?score) ext:TechReportsTextSearch($topic
                                          $n $relevance).
  ?doc TechReports:author ?person
}
```

The text search property function used for searching the technical reports repository again takes three arguments (the search string, the maximum number of results and the minimum relevance score). The search returns the document URI and its

relevance score which are both returned by the SPARQL select query because they provides the evidence for the relevance of the person to the topic. The results of this query will be in the ranked order returned by the single text search which gives a meaningful relevance score and ordering because a single search was used.

For example, this tactic for finding relevant people to a topic returned the groups of (fictious) authors of documents relevant to one of the seed topics.

- David Shaken, Neil Arrested, Ant Fame, Iris Retinex
- Daniel Lyon, Ron Glass, Gary Circle
- Neil Arrested, David Shaken, Ace Beach
- Daniel Lyon, Ron Glass
- Ron Glass
- David Shaken, Neil Arrested, Ant Fame, Iris Retinex
- Matt Goat, Kelvin Chemistry
- Peter Wilder
- Ernest House

Two of the authors (Ron Glass and Peter Wilder) were also authors of cited papers and/or related patents cited in the subpoena. This provides further corroboration of the relevance of these people to the case. The relevance of an entity to the case is corroborated when the same entities are returned by distinct search paths. Several people who were authors of these cited papers or patents did not show up in the search because they did not write any technical reports, but they did write lots of patents and so a similar query on a patents database would also find patents related to the topic.

So we can derive a larger set of people who are potentially related to some of the seed topics without using any of the people seeds given by the subpoena. Not all of these are as good as each other, but the emphasis in e-discovery is on performing a search with high recall and will not be discarded at this stage. Some of these seeds were directly derived from the patent and some were corroborated by the searches. i.e. when the same people were derived using different search strategies.

4.4 Expanding the Set of People

There are several tactics with which we can expand the set of people considered relevant to the case using the structured and unstructured data sources. These are simple tactics that take a person and retrieves a larger set of people using only the structured data sources and not used the hybrid search capability. However these simple tactics can be combined with some other tactics deriving a set of people relevant to a topic which can be used to corroborate or rank the expanded set of people. For example, a simple tactic uses the organizational hierarchy to expand the potential set of people. This exploits the heuristic that people in the same group have similar skills or work on similar products (which is not always true but which is still useful during identification), and so would be likely to possess information relevant to the case.

```
select ?person {
 ?manager hp_org:manages $seed_person.
 ?manager hp_org:manages ?person
}
```

Similarly other tactics use the semi-structured data sources. A simple tactic expands the set of people by retrieving the co-inventors of the patents written by the seed inventor.

```
select ?person {
    ?patent Patents:inventor $inventor.
    ?patent Patents:inventor ?person
}
```

A similar tactic uses the technical report repository to retrieve the co-authors of documents patents written by a given person.

Such tactics can be combined with other tactics which derive people relevant to a topic to obtain stricter queries. For example, the following tactic makes the expansion to all organizational peers of the seed person conditional on the manager having written a relevant patent.

```
select ?person ?doc ?score {
 ?manager hp_org:manages $seed_person.
 ?manager hp_org:manages ?person.
  (?patent ?score) ext:PatentTextSearch($topic
                                        $n $relevance).
  ?report Patents:inventor ?manager
}
```

4.5 Expanding a Topic to Generate Related Topics

The technical reports repository can also be used to derive other related topics. The repository has title and abstract fields for each document that are good sources of the keywords used to characterize a topic. For example the following tactic retrieves the titles of relevant technical reports to create an expanded set of topics.

```
select ?title {
  ?doc ext:TechReportsTextSearch($topic,
                            $n,$relevance).
  ?doc TechReports:title ?title
}
```

We obtained an expanded set of topics (fictious) with one of the initial seed topics.

- “Ink location for Color Halftones”
- “Geometric Screening”
- “Fundamental Characteristics of Halftone Textures”
- “Curved dithering”
- “Multi-pass Error-Diffusion for Color Halftones”
- “Lossless Compression of half-toned Images”
- “Anti-aliasing Methods for Computer Graphics Hardware”
- “Inverse Half-toning for Error Diffusion”
- “n-Simplex gamut mapping”

Some of these topics are inappropriate. Instead of the topics being concerned with printing the topics are concerned with computer graphics and image compression. Some topics whilst concerned with printing are concerned with color gamut mapping and are not relevant to this particular request. Such related topics might still be useful indicators of people with related skills and interests.

4.6 Corroboration

We can corroborate the current set of relevant entities or concepts whenever an existing entity or concept can be retrieved independently by one of these tactics. This was encountered earlier in this paper, when we found that "Ron Glass" was also: an inventor of a related patent, a publisher of a cited paper, and the author of a technical report responsive to one of the seed topics. We can strengthen the constraints on some of our tactics by introducing a requirement for such corroboration.

Similarly if we suspect two different entities or concepts have a relevant relation, we can corroborate this relation by deriving some other common entities starting from either. For example we can corroborate the relation between a topic and product by both using a tactic to find people related to topic, and using another tactic to find people within organizations responsible for this product. The directness of the relationship of people to the product will corroborate the relationship between the topic and the product.

Unfortunately if we perform a topic search of the technical reports repository we find that every author of a responsive document will occur in Labs which is not responsible for any product line. Either we need a different repository, or we need to use a weaker relation between people and products. For a weaker relation we use the existence of an email message relevant to a topic between the person from labs and someone in the business.

The tactic below takes a topic and a product as arguments, plus some controls for the two searches that are used. One performs a search of the technical reports, but the other performs a parameteric search capability on the email repository which restricts the search to emails between two people. The tactic returns the evidence that the topics are related to the product. This evidence is the document and email that are relevant to the topic, and the people who communicated about the topic – one of whom is in an organization responsible for the product.

```
select ?personA ?personB ?doc ?dscore ?email ?escore{
(?doc ?dscore) ext:techreportsTextSearch($topic,$n1,
                                  $relevance1).
 ?doc techreports:author ?personA.
 Sproduct productmaster:product_line ?product_line.
 ?product_line hp_org:is_in_org ?org.
  ?personB hp_org:is_in_org ?org.
 (?email ?escore) email:EmailTextSearch(?personA,?personB,
                          $topic,$n2, $relevance2).
}
```

This evidence could then be analyzed to provide a more sophisticated scoring of the quality of the corroborating link between the topic and the product. We need some

means of scoring the corroboration that would take into account either the quality of the relevance of the technical report or the email message to the topic, or the number of relevant communications between the two people. At the moment we just return the evidence.

5 Conclusions

We implemented a hybrid search and query by extending SPARQL using property functions which returned ranked search results. This gives a rich retrieval model allowing text search and query of structured and semi-structured data to be used together. This was used to exploit product and organizational structure to increase recall by finding potentially related people and products. Furthermore problems with scale are avoided because the structured data sources are used to guide the search and so avoid searching everything in the enterprise.

Unfortunately both product and organizational data sources are constantly changing. For our approach to be most effective there is a need to find the organization and product structure for the particular periods of time relevant to the e-discovery request. This need not be information that is kept during the normal operation of a business. For example, it is common to keep only the current organizational information.

E-discovery legislation only requires an enterprise to disclose information that is held by the enterprise for the normal operation of its business. It does not require enterprises to store additional information. In fact, it is for this reason that proactive records management systems are proposed for e-discovery, as they enforce policies which stipulate that only data with a business purpose should be kept. However maintaining historical organizational and product information to help e-discovery is optional because this is not part of the normal operation of a business. Interestingly the product and organizational structure themselves may not be relevant ESI. They are merely a means of finding relevant ESI and perhaps of understanding its significance.

There are further opportunities for the application of semantic technologies in e-discovery:

- Litigation readiness – where semantic technologies can be used to represent knowledge which helps the finding of ESI without keeping additional ESI. In our investigation we assumed that structured information about people and products was available, which was not in itself ESI but which could help to find ESI which already exists. An important business issue is deciding what information should be kept.
- An information map of the data sources available in the company.
- Representing the semantics of data selected during e-discovery.
- Describing the provenance of data through the different stages of e-discovery.

Acknowledgements

Brian McBride, Dave Reynolds, Mark Butler, and Ian Dickinson were instrumental in developing the approach to e-discovery followed in this paper.

References

1. Rothstein, B.J., Hedges, R.J., Wiggins, E.C.: Managing Discovery of Electronic Information: A Pocket Guide for Judges. Federal Judicial Center (2007), http://www.fjc.gov/public/pdf.nsf/lookup/eldscpkt.pdf/file/eldscpkt.pdf
2. Prud'hommeaux, E., Seaborne, A. (eds.): SPARQL Query Language for RDF. W3C Recommendation (2008)
3. Federal Rules for Civil Procedure (2008), http://www.uscourts.gov/rules/CV2008.pdf
4. Francis, J.C., Schenkier, S.I.: Surviving E-discovery. In: Magistrates Workshop (2006), http://www.fjc.gov/public/pdf.nsf/lookup/MagJ0608.ppt/$file/MagJ0608.ppt
5. TREC Legal Track, http://trec-legal.umiacs.umd.edu/
6. Example Electronic Discovery Request, http://www.fjc.gov/public/pdf.nsf/lookup/ElecDi13.pdf/$file/ElecDi13.pdf
7. Example Preservation Order, http://www.fjc.gov/public/pdf.nsf/lookup/ElecDi21.pdf/$file/ElecDi21.pdf
8. Example Meet and Confer Form, http://www.fjc.gov/public/pdf.nsf/lookup/ElecDi22.rtf/$file/ElecDi22.rtf
9. The Sedona Conference, http://www.thesedonaconference.org
10. Electronic Discovery Reference Model, http://www.edrm.net
11. EDRM Search Guide. EDRM (2009), http://www.edrm.net/files/EDRM-Search-Guide%20v1.14.pdf
12. Baeza-Yates, R., Ribeiro-Neto, B.: Modern Information Retrieval. Addison-Wesley, ACM Press (1999)
13. Chaudhuri, S., Ramakrishnan, R., Weikum, G.: Integrating DB and IR Technologies: What is the Sound of One Hand Clapping? In: Proceedings of the 2005 CIDR Conference (2005)
14. Cafarella, M.J., Re, C., Suiciu, D., Etzioni, O., Banko, M.: Structured Querying of Web Text. In: 3rd Biennial Conference on Innovative Data Systems Research (CIDR), Asilomar, California, USA (2007)
15. Cheng, T., Yan, X., Chang, K.C.: Supporting Entity Search: A Large-Scale Prototype Search Engine. In: ACM SIGMOD 2007, Beijing, China (2007)
16. Goldman, R., Widom, J.: WSQ/DSQ: A Practical Approach for Combined Querying of Databases and the Web. In: ACM SIGMOD International Conference on Management of Data (2000), http://www-db.stanford.edu/~royg/wsqdsq.pdf
17. Hatcher, E., Gospodnetić, O., McCandless, M.: Lucene in Action. Manning Publications (2004) ISBN 1932394281
18. Sesame - the like operator, http://www.openrdf.org/doc/sesame/users/ch06.html#section-like
19. Virtuoso - bif:contains full text search, http://docs.openlinksw.com/virtuoso/rdfsparqlrulefulltext.html

20. Glitter - textlike and textmatch operators, http://www.openanzo.org/projects/openanzo/wiki/SPARQLExtensions
21. Minack, E., Sauermann, L., Grimnes, G., Fluit, C., Broekstra, J.: The Sesame LuceneSail: RDF Queries with Full-text Search. NEPOMUK Technical Report (2008), http://www.dfki.uni-kl.de/~sauermann/papers/Minack+2008.pdf
22. Andreou, A.: Ontologies and Query Expansion. Master of Science, School of Informatics, University of Edinburgh (2005), http://www.inf.ed.ac.uk/publications/thesis/online/IM050335.pdf
23. Crestani, F.: Application of Spreading Activation Techniques in Information Retrieval. Artificial Intelligence Review 11(6), 453–482 (1997)
24. ARQ home page, http://jena.sf.net/ARQ
25. Carroll, J.J., Dickinson, I., Dollin, C., Reynolds, D., Seaborne, A., Wilkinson, K.: Jena: Implementing the Semantic Web Recommendations. In: Proceedings of the 13th International World Wide Web Conference (2004)