RC24420 (W0801-118) January 29, 2008
Computer Science

IBM Research Report

Improving Information Access for a Community of Practice Using Business Process as Context

Yu Deng, Murthy Devarakonda, Ruchi Mahindru, Nithya Rajamani, Norbert Vogl, Wlodek Zadrozny
IBM Research Division
Thomas J. Watson Research Center
P.O. Box 704
Yorktown Heights, NY 10598

Research Division
Almaden - Austin - Beijing - Cambridge - Haifa - India - T. J. Watson - Tokyo - Zurich

Improving Information Access for a Community of Practice Using Business Process as Context

Yu Deng, Murthy Devarakonda, Ruchi Mahindru, Nithya Rajamani, Norbert Vogl, Wlodek Zadrozny
IBM T.J. Watson Research Center
P.O. Box 704
Yorktown Heights, NY 10598

Abstract

This paper addresses the important problem of finding relevant information in the context of a business process. It presents a solution called EIL (Enterprise Information Leverage) which combines information extraction and semantic search techniques to support information needs of IT Services sales professionals. Structured and unstructured data is leveraged using a novel architecture and specialized search algorithms. EIL organizes information around business activities (e.g. a sale) and supports semantic concept-based information retrieval by utilizing a combination of directed database queries and document searches where the relevant business activities act as a contextual constraint. Experiments show that this approach is effective for reducing noise in search results and for providing useful business context. Production rollout of EIL is currently underway within IBM.

1 Introduction

A recent IDC [9] study on costs incurred by enterprises associated with information work shows that professionals spend roughly 36% of their time on collaboration through emails, about 24% of their time for search and another 24% for analysis of the data. Therefore solutions that help professionals with their information needs in the business context have the potential to make a significant impact on the organization's bottom-line. The study also emphasizes the need for social networking for exchange of knowledge and collaboration in organizations. More importantly, as the study has concluded, knowledge work is a complex mixture of intellectual work and repetitive tasks. Hence search solutions that just provide a search-box experience do not help these professionals. With the economy shifting more towards information-based knowledge enterprises, it becomes critical to address this gap.

In this paper, we address the problem of finding relevant information within the context of a business process[1] for a community of practice. Specifically, the community of sales professionals in the IT Services business who have the responsibility to generate revenue by winning sales opportunities. Their business process, which is the process of selling IT services, calls for significant and specific information needs. An important point to observe is that sales professionals think of information in a context of their business activities (i.e. instances of business processes). This led to the design of Enterprise Information Leverage (EIL), an information access solution that this community would like to have.

EIL significantly deviates from typical "one-size-fits-all" enterprise search initiatives in that its search leverages business process as a context. The methodology enabling the EIL search solution requires an understanding of the information needs of sales executives in the context of their selling process and the tasks they face; and an understanding of the information sources (both structured and unstructured) and the content quality of the sources. It then applies text analysis techniques to enable semantic search capabilities and primes them with business related heuristics. As one can see, every step in our methodology makes a connection between the business, its users and the technology that enables search and discovery.

Our contribution lies in developing and evaluating a general methodology for building an information access solution where the information needs in a business context are used to constrain the answer sets and ensure high quality results. Specifically, we

- Created an architecture and algorithms for an information access system focused on information needs restricted to a business process, taking access control and confidentiality into consideration.
- Integrated innovative text analytics with semantic search and database access to provide up to date an-

[1] A business process is a collection of interrelated tasks, which solve a particular issue. (from http://en.wikipedia.org/)

swers in the business context.

- Investigated and classified information needs within the process of selling IT services.

- Built a system embodying this architecture and reflecting the specific information needs of IT Services sales personnel.

- Evaluated the system with beta users (with a very positive feedback) and by analyzing a collection of queries and results. The evaluation helps us understand how EIL compares to a business-agnostic search-box kind of search.

- Put the system into production environment for rollout.

- Drafted a plan for improvement of the system as more data becomes available and additional evaluation is performed.

Technically, our solution combines text analysis with database queries and document searches. This is complemented by access controls so that authentication and authorization are a part of the solution. The front-end is oriented towards supporting the relevant business process and a sales executive can issue a query on certain criteria (concepts) in addition to performing keyword searches. The returned results are different from the typical "search-box" approach. What comes back is a set of relevant business activities, along with the context information about them (which satisfies the repetitive information requirements of sales executives). if access permissions allow, a set of relevant documents that support the extracted context and the query is also returned.

The architecture of the system is described in [21], which highlights the business rational behind the project of EIL, presents the details of our study on the information needs of the community of practice and discusses the EIL implementation at a high-level.

In this paper we describe three steps of our approach, which can be generalized for similar business processes:

1. Understanding information needs of sales executives in the context of their business processes and the challenges they face; and analyzing our end-users' mental information roadmap to identify a set of important concepts (Section 2).

2. Leveraging text analysis techniques to enable business-driven semantic search capabilities and priming them with business related heuristics. This involves creating an architecture, developing text annotators, and ensuring access control to confidential information (Section 3).

3. Evaluating the system to understand how it compares to a business-agnostic search-box kind of search within the context of the specific information needs (Section 4).

In addition, in Section 5, we place our work within the context of prior results in enterprise search, semantic analysis and information retrieval from databases.

2 Representing Information Needs of Business Professionals through Meta-queries

To get insights on the information needs of the target community of practice (Sales Executives), we analyzed their communications in an email distribution list. The distribution list is meant for them to collaborate and share knowledge. It typically has queries soliciting response for information needs associated with business activities they are engaged in. We monitored and analyzed 120 email threads (accumulated over period of nine months). The details of the study, together with our findings on interviewing them are documented in [21]. Key findings include:

- The lack of effective search and discovery tools available to sales executives which provide answers to their business questions. Sales executives expect responses from a search system that reflects their business needs than a document search infrastructure.

- People identification is one of the key information tasks in the sales community. It is crucial to find the right person who is involved in a similar situation or can share tacit and further explicit knowledge. Therefore, social networking is very important. For example, including both explicit and implicit solicitation, 63 out of the 120 email discussions ask for social networking information. This confirms that search is not the "end" but only a means for information access. The study reports in [19, 17] share the same insight.

- The sales professionals have focused information needs that are sufficiently constrained by their job role and business process and hence different from those of consumer search. But simultaneously, their information needs are also complex in nature. So if their query were to be answered by issuing keyword searches, it would take several attempts and significant time to process the results and deduce the information they are looking for.

From our study, we derived four key meta-queries which represent typical queries in the email distribution list [2]:

[2] Sometimes they are an inherent part of a larger query instead of a standalone query by themselves.

- Which business engagements have a scope that involves <this service>?
  Approximately 38% of emails are of this type. The scope of a business activity is often an important criterion to select activities of business interest.

- Who in <this role> has worked with <this person> in <this organization>?
  Approximately 17% of emails are of this type. The selling process is intensely client-relationship oriented and the sales executives need to be aware of all the different connections and linkages from the organization into the client.

- Who has worked in the capacity of <this role>?
  About 36% of emails fall in this category. Sales community leverages expertise of different other communities of practice in order to execute well in their job role. They prefer to talk to the subject matter experts instead of poring over the details of explicitly created information even when it is available and accessible.

- Who has worked on <this service> that involved <this keyword>?
  About 29% of emails are of this type. Not all information needs can be captured by explicit concepts. We realize the importance of search-box functionality to let the user have complete control of the search when desired.

Our business-activity driven search helps the sales community find answers to the kind of queries discussed above. It is important to note that these are the most typical of supported queries. We also chose them for the purpose of evaluation (See Section 4).

3 Business-Activity Driven Search

The methodology of business-activity driven search has two critical elements. First, an end user's business information needs can be resolved into a set of semantic concepts *that are relevant to the user's role* in the business process, and the user may search the data using these concepts in addition to general key word search. For example, win strategy is one of the concepts of importance to a sales practitioner. Therefore, identification of important concepts to be presented for a user community is a part of the methodology. Second, a search query *returns a set of the most relevant business activities first* rather than documents or links. A synopsis context is provided for each business activity and the user may further explore most relevant documents within a business activity based on its synopsis. The methodology therefore enables its users to identify the most relevant business activities, provides access to the business context and people involved in the activity as a part of the synopsis, and relevant documents from that activity as needed. Security and privacy concerns limit what a user can see based on their role and access controls on the data.

```
BUSINESS-ACTIVITY DRIVEN SEARCH
/* Input: A query on semantic concepts and keywords */
/* Output: A set of relevant business activity synopses and (upon access
     permissions) a set of relevant documents for each activity */
1)  Identify fields with user-entered values
2)  Compose synopsis query from Form-based input
3)  Compose SIAPI query from Form-based input
4)  Execute synopsis query
5)  if the result set of synopsis query is not empty
6)      S ← business activities returned by synopsis query
7)      if SIAPI query is not null
8)          Execute SIAPI query against the semantic index for only
            the activities in S
9)          R ← business activities returned by SIAPI query
10)     else
11)         R ← S
12) else
13)     if SIAPI query is not null
14)         Execute unscoped SIAPI query
15)         R ← business activities returned by SIAPI query
16)     else
17)         R ← empty set
18) Rank the results in R
19) Present the results to user with proper access control
```

Figure 1. Major Steps and Information Flow of Business-Activity Driven Search

This methodology is supported by our search algorithm. In Figure 1, we highlight the major steps and information flow of business activity driven search. For example, steps 5 through 7 are used to restrict results to relevant business activities. Notice that the search is supported by two parts: synopsis search against the business activity context in DB2 database and SIAPI (Search and Index API) search against the semantic index in OmniFind [2]. In step 17 of Figure 1, the results are ranked by taking both the ranking from synopsis query and the ranking from SIAPI query into account. Specifically, we normalize the document relevance scores from OmniFind (e.g., compute an average score) and then combine the normalized score with the synopsis relevance score.

3.1 System Architecture

Figure 2 shows the architecture of the proposed information retrieval solution called *Enterprise Information Leverage* (EIL). The solution is based on text analytics to automatically annotate data and documents with relevant semantic concepts and extracting valuable domain-specific information (e.g., list of key contact persons involved in a

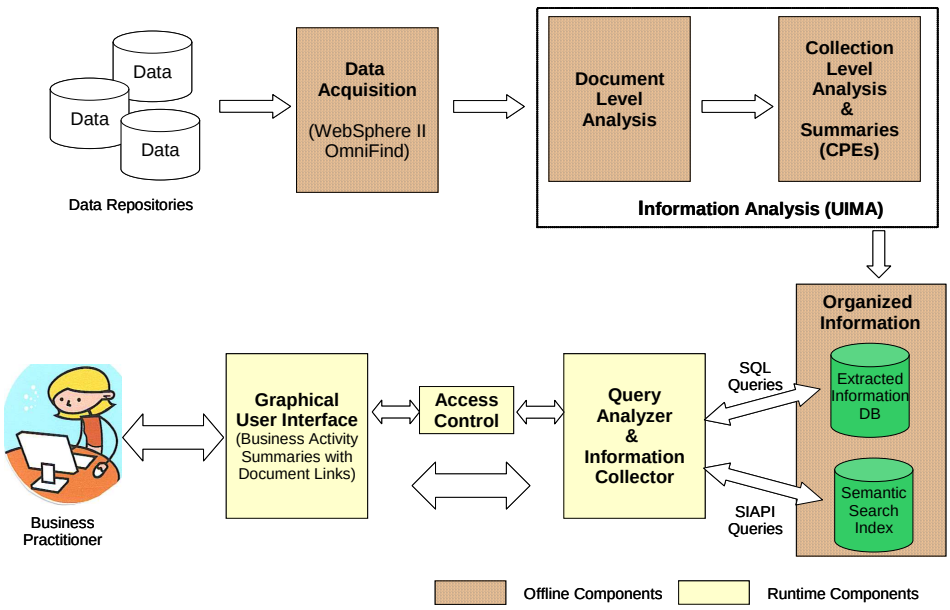

Figure 2. The Enterprise Information Leverage (EIL) Architecture

business activity) from the annotated corpora. Once this information is extracted, it is integrated and stored in a structured database as part of the business context.

EIL achieves the aforementioned by using *offline* and *online* (see differently shaded parts of Figure 2) components. The offline components are responsible for crawling various data repositories (**Data Acquisition**), extracting the relevant information from the data (**Information Analysis**) and putting them in a structured format (**Organized Information**). The online components are responsible for retrievng results in response to a user's query and presenting them to the user in a rank order. Specifically, the **Graphical User Interface** (Lotus Notes based) takes a user's query (about interested concepts and keywords), and sends it to the **Query Analyzer** & **Information Collector** component which converts the query into a SQL query for the DB2 database and (possibly) a SIAPI (Search and Index API) query for the semantic index. Meanwhile, the **Access Control** component takes care of security and confidentiality. For example, if a user is not authorized to access a data repository, the system presents to the user only a synopsis of the desired information including a list of contact persons with whom the user could communicate. More details on architecture can be found in [21].

3.2 Developing EIL Annotators

The information needs of sales professionals, obtained from our study mentioned earlier, are captured in a manageable number of semantic concepts. EIL needs to support search based on these concepts. We developed a number of annotators (text miners) to support such a concept-based search.

We provide an illustration of an annotator in Section 3.2.1 and general guidelines derived from our experience in Section 3.2.2.

3.2.1 Social Networking Annotator

As shown in Section 2, social networking is a key aspect of our business practitioners' daily work. In Figure 3, we present the details of social networking annotator, which extracts the information of key contact persons in a deal.

In the algorithm, an example of step 6 could be inferring the name and organization details of a person from their internet email address where the pattern is firstname.lastname@organization.com. This ensures that more structured fields are available for structured queries and further analysis. Meanwhile, the reason for step 10 is that there may be several entries for the same person and we need to merge the different fields into one single record or alternatively use document information and metadata (e.g., business importance of the document) to determine the relative priorities and assist selection between conflicting values of fields.

Though each data repository we are dealing with has a pre-defined template for entering the names, roles and contact information of the people who worked on a particular business activity, often this is not populated or properly maintained. Also the roles get defined only after the initial kick-off and are reflected in semi-structured forms in scattered places in the data repositories. These issues severely

```
SOCIAL NETWORKING ANNOTATOR
/* Input: a collection of unstructured and semi-structured
        documents */
/* Output: social networking annotation (including name, email,
        phone, organization, role etc) */
1)  I ← the subset of documents identified to be candidates
2)  E ← the subset of documents to be excluded irrespective of I
3)  for each document in I but not in E
4)      Identify the business activity of this document (metadata)
5)      Process the document text and metadata
6)      Infer values of other related fields from existing fields.
7)  end for
8)  Write annotations into a roll-up file for collection-level
    processing.
9)  for each business activity in the collection
10)     de-duplicate social networking annotations in the rollup
11) end for
12) Normalize the fields to remove semantic ambiguity.
13) Collect personnel information from intranet repositories
    to update the annotations
14) Populate the annotations into a set of tables in DB2 database
    as part of the corresponding business context
```

Figure 3. Algorithm for Social Networking Annotator

affect the quality of annotations. We leverage the personnel information from the intranet to improve the quality of extracted information (see step 13 in Figure 3).

In addition, we leveraged business related heuristics and took advantage of the structure of the documents when available. We could further leverage machine learning techniques to help us identify the candidates for the annotator in order to improve the quality. An alternative to the above approach would be to use advanced entity analytics to identify names and use patterns to annotate phone numbers, emails etc., and then use co-occurrence techniques to connect them up. For example, if the business activity typically involves creation of a spreadsheet with the team members' information, leveraging the process conventions on the title/headers and semi-structured format (rows and cells) that can help with the information extraction would perform better than just blindly applying patterns interpreting the entire data as a blob of text.

Here we point out two side effects of the semantic analysis in EIL. First, it changes the way search queries are issued in that a query in EIL is always specified within a context (e.g., within certain industry or having certain scopes[3]). This is because the results of semantic analysis are populated into the database as part of the corresponding business context, as shown in step 14 of Figure 3. Second, it changes the information to be presented such that the information is synthesized from both extracted details from documents and auxiliary data from reliable data sources. This is due to the step where the analysis engines integrate data from multiple sources to improve the quality of annotations.

[3]The scope of a business activity defines specific roles and responsibilities. For example, "Computer operations and monitoring" could be an item in the scope of IT services.

3.2.2 Guidelines for Building Annotators in EIL

For each type of EIL annotator, in Table 1, we list the applicable data sets, any required preprocessing steps, advantages, limitations and suggestions for improvement. These guidelines are based on the issues we have encountered as well as our development experience. We believe that they are valuable for annotator developers.

In Table 1, the regular expression-based annotator is the easiest one to implement, but it can only express simple patterns of data sets. Cleansing the data sets and applying domain knowledge to the regular expressions can help retrieve more accurate results. On the other hand, adopting ad-hoc heuristics in the annotator can also improve the accuracy. Such heuristics, however, are highly dependent on the data sets and it is time-consuming for developers to identify them. Notice that both regular expression-based and heuristics-based annotators cannot capture data semantics, e.g., relationships and constraints, which can be covered by an ontology-based annotator. The performance of an ontology-based annotator, however, highly relies on the quality of ontologies. Therefore, refining the associated ontologies is important for this type of annotators. The four types of primitive annotators may be assembled into a composite one, capturing complicated control and data flows.

The semantic analysis techniques for each of the aforementioned annotators are not new. But a single one of them is not enough to satisfy the requirements of the EIL system. It is their combination that matters, because only the right combination produces the desired results.

3.3 Using Structured Information

We have learned that leveraging structured information wherever possible in annotators improves the quality of information extraction. Below are the highlights of our experience.

Custom Parsing: Semantics can be captured and extracted from document structure and conventions. For example, a PowerPoint presenter uses title and subtitle to convey the key point that he would like to cover. It is important to preserve the structure of documents during the parsing phase so that our annotators can make use of it in the phase of information analysis (see Figure 2). We have developed components to annotate and tag the structure of PowerPoint presentations and Excel spreadsheets.

Data Integration: Recently, there has been an increased interest in regarding Web as a set of inter-connected

Annotator Type	Applicable Data Sets	Preprocessing Step	Advantages	Limitations	Suggestions for Improvements
Regular Expression-based	Any	Familiarity with patterns of the data sets	Simple; easy to implement	Limited expressiveness	(1) Cleansing the data sets; (2) Applying domain knowledge to refine the regular expr.
Heuristics-based	Any	Familiarity with the domain	Quickly identifying relevant pieces of information	Ad-hoc; Highly dependent on the data sets	(Semi-)automatically assisting in identifying heuristics
Ontology-based	Ontology associated data sets	Building/reusing existing ontologies	Capturing data semantics in the annotator	Highly dependent on the quality of ontology	Iteratively refining the ontology with the output of annotator
Classifier-based	Any	Building the classifier	Capturing complex & abstract concepts	Highly dependent on the training data set	Enhancing classifier with semantic information or ontologies
Composite	Any	Specifying data and control flow	Capturing complex control and data flows	Manual composition	(Semi-)automatically identifying relevant primitive annotators and and assemble them

Table 1. Guidelines for Developing Annotators in EIL

databases, not a set of documents, for building a high-quality search engine [24]. This view is also applicable in an enterprise intranet. For example, the internal personnel website has a hidden database containing each employee's information. The data sources are structured and have well-defined semantics. We then integrated information from these data sources to validate and update the analysis results from our annotators. An example is the social networking annotator, through which we integrated data from our internal personnel website to validate the extracted people's status and update their contact information.

3.4 Collection Processing Analysis

As mentioned earlier, the results of the document analysis done by the annotators are processed by the Collection Processing Engines or CPEs (see Figure 2; Information Analysis). These engines can be modeled to execute various reasoning techniques to verify whether the results produced by the annotators are correct, as well as to infer some new collective results. The collection processing step is important for enabling EIL methodology. For instance, scopes of business activities are first extracted by a document-level annotator and then fed into a CPE, which aggregates them across a business activity, counts their occurrences with regard to the activity and identifies the ones that can be regarded as its scopes. Furthermore, the CPEs can also be designed for doing multiple post-analysis tasks on the results obtained from the annotators such as removal or normalization of duplicate/redundant data.

4 Search Quality Analysis

In this section, we show the improvement in search quality by deploying EIL to search engagement workbooks[4] and present a detailed analysis of queries corresponding to the four typical meta-queries (see Section 2) from the email distribution list of the sales community. We compare the way queries are framed and the results obtained through EIL vs keyword search in OmniFind [2].

The engagement workbooks we used for the experiment contains approximately about 15,000 documents in total corresponding to 23 IT Services activities (we call a services activity an engagement or a deal). Since the engagements on which these experiments were carried out were real, we have anonymized them to protect the confidential information. The below experiments also assume that there are no access controls on the documents for any sales professional issuing the queries. The information gain will not be the same with keyword search when access control restrictions are in place for any random document access by users, whereas EIL supports focused information access by extracting information which can be regularized and hence does not face the same constraints.

4.1 Result Analysis

Meta-query 1: Let us take the sample query - Which business engagements have a scope that includes End User Services? A simple keyword (OmniFind) search based on the keyword End User Services or EUS returned 261 documents. Apparently, the query had overlooked that End User Services has two subtypes: Customer Services Center and Distributed Computing Services. With the sub-

[4] An engagement workbook is a data repository containing documents related to IT services.

types explicitly considered, OmniFind returned 1132 documents, as shown in Figure 4. Again, to find out relevant deals, a user has to read through some or most of the returned documents. Just a mention of CSC (Customer Services Center) in any document would not mean that it is a part of the engagement scope. Neither is the phrase "CSC" used to describe the service consistently throughout the organization.

EIL supports concept-based search, by which a user can specify one or more services scopes to locate relevant deals. EIL exposes the services in scope for the engagements based on analyzing different business documents. The documents that are candidates for having services information in them are identified and analyzed to annotate the different services - it leverages a simple taxonomy for performing the annotation. Then all annotations in an engagement are analyzed to come up with the services that have the most significance in the engagement. Hence for the query about End User Services, EIL searches the deal synopses and returns a list of relevant deals, as shown in Figure 5. The order of the services (called towers in the diagram) reflects the relative significance of the towers. So the deal C here is primarily a Customer Service Center based engagement.

Furthermore, a user can see the entire business context of a deal by clicking on the deal name. Figure 6 presents the contextual information of deal C, e.g., customer name, total contract value, contract term, deal team, win strategy and technology solutions, etc.

We have executed 10 similar queries on a set of 12 deals and validated the results of OmniFind keyword search vs EIL by a domain expert. The comparison details (see Table 2) show that although OmniFind keyword search has 100% recalls[5] for all of the queries, it sacrificed on precision[6] and returned plenty of useless information. According to the F-Measure,[7] EIL has higher quality than keyword search, thereby reducing noise in the search results.

Meta-query 2: The sample query here is - Which CSE (Client Solution Executive) has worked with Sam White from company ABC? We first tried the following keywords in OmniFind: Sam White ABC CSE. Nothing was returned. Then we tried Sam White ABC instead, and four documents were returned, from which we were able to find out that Sam White had worked on the business deal called ABC Online. We then searched with the newly learnt deal name ABC Online and the position title CSE, and 97 documents were returned, as shown in Figure 7. By going through a portion of the 97 documents, one could infer that a couple of sales professionals have worked in the deal as a CSE.

[5]Recall is defined as the number of correct answers returned divided by the total number of correct answers that should have been returned.

[6]Precision is defined as the number of correct answers returned divided by the number of answers returned.

[7]F-Measure is defined as $\frac{2*Precision*Recall}{Precision+Recall}$.

Query	Precision		Recall		F-Measure	
	EIL	KW	EIL	KW	EIL	KW
1	0.82	0.75	1	1	0.9	0.86
2	0.38	0.23	1	1	0.55	0.37
3	0.5	0.33	0.75	1	0.6	0.5
4	0.33	0.08	1	1	0.5	0.15
5	0.6	0.5	1	1	0.75	0.67
6	0.55	0.5	1	1	0.71	0.67
7	0.9	0.75	1	1	0.95	0.86
8	0.4	0.5	0.33	1	0.36	0.67
9	1	0.83	0.8	1	0.89	0.91
10	0.33	0.17	0.5	1	0.4	0.29

Table 2. The quality of EIL search vs OmniFind keyword (KW) search

EIL has a more effective solution to address these types of queries. Since it leverages a social networking annotator (Section 3.2.1), it creates an automatic contact list for an engagement by analyzing all the business documents in the deal. Hence we simply entered Sam White and ABC in the *people* search section. By searching the contact list in deal synopses, the system found the deal ABC Online in which Sam White had been involved. The deal synopsis has a tab named *People* that show all the personnel involved in the deal, as well as their titles, organizations, role responsibilities, email addresses, and contact phone numbers. Furthermore, the tab organizes the contact information into several categories based on the normalized role information, which can further help navigation and location of details quickly and conveniently. These categories include core deal team, technical support team, delivery team, client team, third party consultant, etc. Note that while the keyword search could return the names of the CSEs (after a three step querying and an effort to read nearly 100 documents), it did not return other useful information as EIL did. Nor did the keyword search provide the rest of the business context to provide a holistic view of the deal.

While the benefits of using EIL in the above type of queries are hard to measure by traditional evaluation metrics, we argue that it can consistently outperform keyword search when used in the context of sufficiently constrained business activities that require a comprehensive context instead of what a single document can provide.

Meta-query 3: The sample query here is - Who has worked in the capacity of cross tower TSA in engagements (this is compounded with other meta-queries typically, but here we evaluate this in isolation). Fetched with key words, OmniFind returned 149 documents, many of which do not contain anything relevant to the query submitted. The reason is that cross tower TSA happens to be a field name in a certain kind of documents in a deal (an application that records service details to be delivered in an engagement has

Figure 4. (Meta-query 1) OmniFind Search Results for End User Services

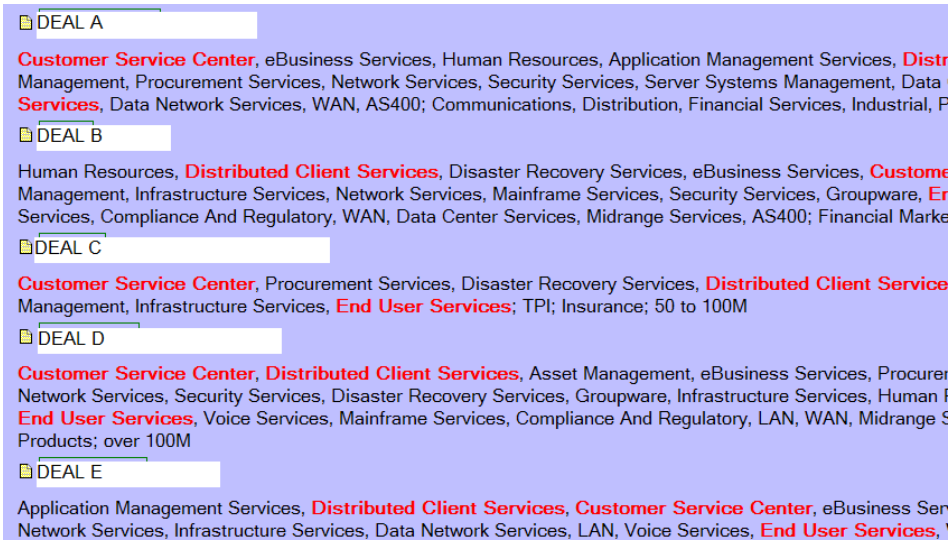

Figure 5. (Meta-query 1) EIL search results for EUS

Figure 6. (Meta-query 1) The synopsis of Deal C

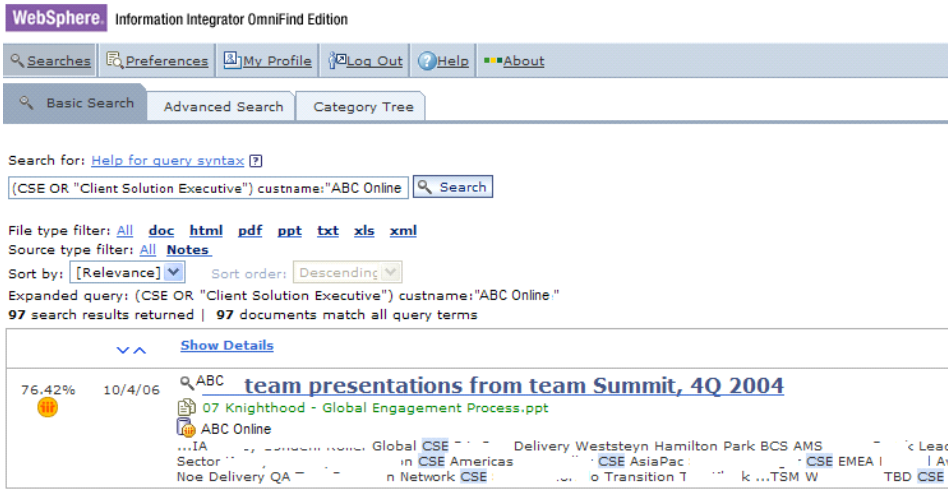

Figure 7. (Meta-query 2) OmniFind search results for CSE of ABC Online

cross tower TSA as a part of its schema) and OmniFind returns the document as a hit even though no value accompanies the cross tower TSA field. Hence it is not worth spending time on poring through the 149 documents to find out that a couple of these (somewhere down the list) have values for this field, which might include only a person's name. Then the user issuing the query would need to go to the corporate repository to issue a second search in order to find all the contact details. But again there is no context about on what kind of deals or with whom this person has worked before.

With EIL, the same query could be run as a search on the entire deal synopsis or only the contact list created from social networking annotator, which quickly locates the contact information and engagements involving the relevant personnel. Due to confidentiality constraints, we do not provide detailed screenshots here.

Meta-query 4: The sample query here is - Who within the organization have done engagements involving data replication technology within the scope of Storage Management Services. A keyword based search for this information would involve multiple steps that combine some of our previous meta-query discussions. The first step would be to issue a search on all the words combined in one shot and analyzing the results. The engagement that satisfies this would be apparent soon if the words happen to be in the same document. If that is not the case, then a part of the search needs to be issued first, i.e. just Storage Management Services, then the engagement is identified and then the next search i.e. data replication, is issued together with the engagement context identified in the previous step. Then after this, many more searches need to be issued with the engagement context to identify the roles that worked on the relevant one. Users can specify the query in EIL as illustrated in Figure 8. The query can be split up into services in scope concept-based search and keyword search in the interface. It could be issued as a keyword search against the entire deal synopsis or only the technology solution overview section or also against the entire engagement workbooks. Since the technology solutions tab in EIL presents the technical solution strategy and overview for the towers in an engagement, the first preference would be to issue searches against those. Another advantage this search brings is that, it is not document search (which would be bounded by access controls).

EIL returned search results are shown in Figure 9 for the Meta-query. Notice that EIL first returns relevant business activities, then for each activity, it lists related documents. In EIL, users can go to the *People* tab of the deals that seem interesting and relevant to identify employees to connect up with. This is made possible by the combination of tower (services in scope) concept, technology solutions overview concept, social networking, together with keyword search capability.

Currently EIL is under production rollout within IBM, where more than half a million documents from almost 1000 engagements have been incorporated into the system. Our user community has provided very positive feedback and some suggestions for improvement on search as well as annotator aspects. This together with a detailed user study has been incorporated as a part of our future work.

5 Related Work

The related work lies in three areas: enterprise search, semantic analytics, and DB/IR. We highlight the key differences as follows.

Most of the enterprise search providers have started integrating search functionality with a customer's specific business processes and information gathering applications [18].

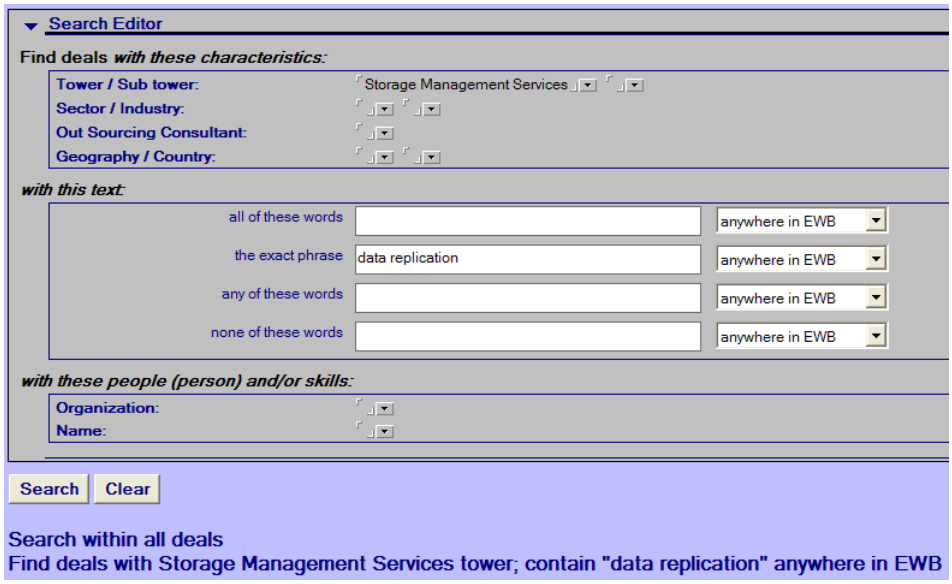

Figure 8. (Meta-query 4) EIL search interface for data replication and Storage Management Services

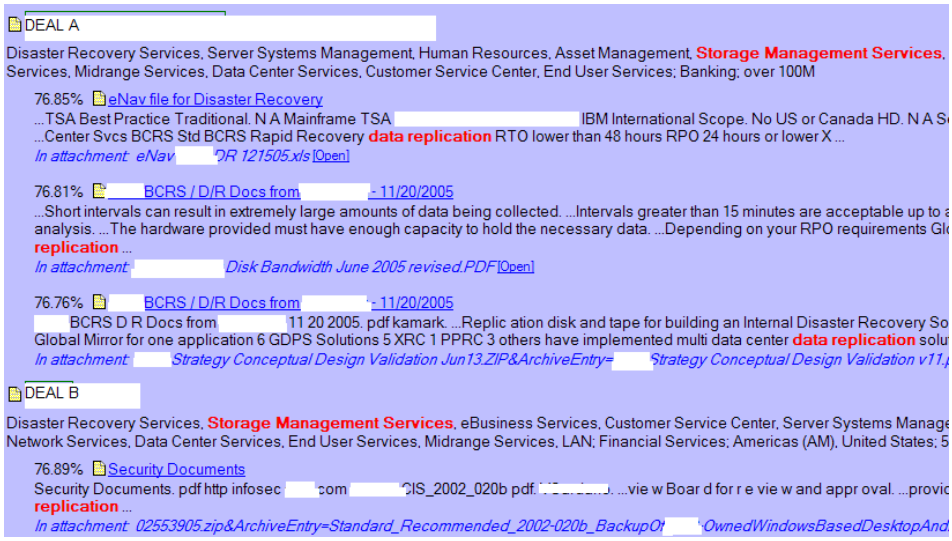

Figure 9. (Meta-query 4) EIL search results for data replication and Storage Management Services

The leaders in this space (Autonomy, Endeca, Fast, Google, IBM, Microsoft and others) have started to use advanced pattern matching techniques, information theory, faceted search, and other techniques [3, 7, 8, 11, 2, 22]. However, our approach (built on the existing OmniFind platform) is different: we more than acknowledge the end-user's view of business processes; we emphasize it to the degree that we have designed the architecture and algorithms to reflect it, as shown in Section 2 and Section 3.

There has also been an emergence of research interest in the field of semantic analytics [6, 20, 16, 12, 5, 4]. However, our work is agnostic with respect to the text analysis techniques (see Section 3.2.2). In current system, we have leveraged UIMA [10] as the text analysis framework, but we are open to other technology.

Several groups have been working on combing database and search technologies to improve the system performance. For instance, [1] and [15] support keyword search in relational database systems, [14] adapts IR-style document-relevance ranking strategies to the problem of processing free-form keyword queries over RDBMSs, and [23] presents similarity metrics to improve relevance ranking in XML databases. In addition, [24] has leveraged deep Web to build a search engine with high quality.

The EIL work does not belong to either one of the categories. It is a search platform designed based on business processes to serve professionals. It has similar challenges with enterprise search systems [13, 19]. In addition, it leverages semantic analytics and DB/IR techniques for delivering a good solution to users. Our research (see evaluation in Section 4) confirms that leveraging structured and clearly-defined data whenever possible improves the search quality and user satisfaction.

6 Summary

We described EIL, a novel information access system for a professional community. We presented motivation, design, and evaluation of the system (currently under deployment). The methodology, architecture and algorithms presented here are applicable in situations where a business process constraints information needs to a limited number of templates/concepts/meta-queries. This approach is general, and similar solutions can be built on a variety of search and database platforms, leveraging different forms of text analytics and enterprise data integration.

Acknowledgment The authors are thankful to Howard Sachar and James Merritt for supporting the EIL production rollout and their continued guidance in making this research useful for practitioners. The authors also want to thank Heather Smith, Joseph Wells and Ray Rose for their system development efforts as well as thoughtful comments and discussions.

References

[1] S. Agrawal, S. Chaudhuri, and G. Das. DBXplorer: A system for keyword-based search over relational databases. In *ICDE*, Washington, DC, USA, 2002.

[2] N. Alur, T. Brown, C. Delgado, R. Isaacs, and M. Przepiorka. *WebSphere Information Integrator OmniFind Edition: Fast Track Implementation*. IBM Redbooks, 2006.

[3] Autonomy technology white paper. http://www.autonomy.com/content/downloads.

[4] P. Buitellar and S. Ramaka. Unsupervised Ontology based Semantic Tagging for Knowledge Markup. In *ICML Workshop on Learning in Web Search*, 2005.

[5] P. Cimiano, G. Ladwig, and S. Staab. Gimme' the Context: Context-Driven Automatic Semantic Annotation with C-PANKOW. In *14th international conference on World Wide Web*, pages 332–341. ACM Press, 2005.

[6] S. Dill, N. Eiron, and et al. SemTag and Seeker: Bootstrapping the Semantic Web via Automated Semantic Annotation. In *12th international conference on World Wide Web*, pages 178–186. ACM Press, 2003.

[7] Endeca. http://endeca.com.

[8] FAST. http://www.fastsearch.com.

[9] S. Feldman, J. Duhl, J. R. Marobella, and A. Crawford. The hidden costs of information work. *IDC White Paper*, Mar. 2005.

[10] D. Ferrucci and A. Lally. UIMA: An Architectural Approach to Unstructured Information Processing in the Corporate Research Environment. *Natural Language Engineering*, 10(3–4):327–348, 2004.

[11] Google Search Appliance. http://www.google.com/enterprise/gsa.

[12] S. Handschuh, S. Staab, and F. Ciravegna. S-CREAM - Semi-automatic CREAtion of Metadata. In *13th International Conference on Knowledge Engineering and Knowledge Management*, pages 358–372. Springer-Verlag, 2002.

[13] D. Hawkings. Challenges in Enterprise Search. In *15th Australasian Database Conference*, pages 15–24. Australian Computer Society, 2001.

[14] V. Hristidis, L. Gravano, and Y. Papakonstantinou. Efficient IR-style keyword search over relational databases. In *VLDB*, Berlin, Germany, 2003.

[15] V. Hristidis and Y. Papakonstantinou. DISCOVER: Keyword search in relational databases. In *VLDB*, Hong Kong, China, 2002.

[16] T. Jayram, R. Krishnamurthy, S. Raghavan, J. Thathachar, S. Vaithyanathan, and H. Zhu. AVATAR Information Extraction System. *IEEE Data Engineering Bulletin, Special Issue on Probabilistic Databases*, 29(1):40–48, 2006.

[17] R. Lewis. Zen and the art of enterprise search. In *CIO Magazine*, `http://www.cio.com/weighin/column.html?CID=24882`, September 2006.

[18] A. Moore. Best practices in enterprise search. *KM-World Magazine*, May 2007.

[19] R. Mukherjee and J. Mao. Enterprise Search: Tough Stuff. *ACM Queue*, 2(2):36–46, 2004.

[20] B. Popov, A. Kiryakov, D. Ognyanoff, D. Manov, and A. Kirilov. KIM – A Semantic Annotation Platform for Information Extraction and Retrieval. *Natural Language Engineering*, 10(3–4):375–392, 2004.

[21] N. Rajamani, M. Devarakonda, Y. Deng, W. Zadrozny, and J. Pathak. Business-activity driven search: Addressing the information needs of services professionals. In *International Conference on Services Computing*, Utah, USA, 2007.

[22] Microsoft office sharepoint server 2007. http://office.microsoft.com/sharepoint.

[23] A. Theobald and G. Weikum. The index-based XXL search engine for querying XML data with relevance ranking. In *EDBT*, pages 477–495, Prague, Czech Republic, 2002.

[24] Z. Zhang, B. He, and K. C.-C. Chang. Understanding web query interfaces: Best-effort parsing with hidden syntax. In *SIGMOD Conference*, pages 107–118, Paris, France, 2004.