

Business-Activity Driven Search: Addressing the Information Needs of Services Professionals

Nithya Rajamani¹, Murthy Devarakonda¹, Yu Deng¹, Wlodek Zadrozny¹, Jyotishman Pathak²

¹IBM T.J. Watson Research Center, P.O. Box 704, Yorktown Heights, NY 10598

²Department of Computer Science, Iowa State University, Ames, IA 50011-1040

{rnithya, mdev, dengy, wlodz} AT us dot ibm dot com, jpathak at cs dot iastate dot edu

Abstract

Services business generates significant amount of human and machine created data. Search and discovery of relevant information from this data is a critical factor in enhancing workforce efficiency and operational excellence in such a business. The challenge is how to leverage the interplay among information retrieval techniques, role of human experts, and business processes. This paper presents a case for using “business activity” (e.g. a sale) as a contextual basis to meet the challenge. We developed a tool based on this approach to address the needs of services professionals. The tool is designed to support role-specific concept based queries that return the most relevant business activities first and subsequently enable retrieval of the most relevant documents and information for the business activity. Any relevant knowledge privy to an expert can be obtained by first finding the most relevant sale, followed by corresponding participant information, and finally by communicating with the participant expert. Business activity as a critical context enhances the search quality while bringing in the social network and business process. This approach suggests a new methodology for search and information management that delivers better business results for the services business.

1. Introduction

It has become increasingly critical to leverage information insights for better decision making in a business process. This is especially true in the services business because of the intrinsic complexity of an information based enterprise and the unique value customization can bring to a customer. In practice, the services business involves a confluence of experts from different domains who have different responsibilities and information needs. For example, an attorney on the team may be concerned about reducing the risks from legal liabilities whereas the sales executive is focused on closing the most

profitable business deal. In the end, a team of professionals drive the business process and they want to succeed together. Thus, it is necessary to provide tools that not only focus on the information needs of the practitioners but also strengthen the underlying services processes and collaboration.

In order to better understand the information needs of a community of practitioners in services, we studied the email distribution list of a services business sales community, which is used by the sales professionals to seek information from their counterparts. We analyzed approximately 120 discussion threads that accumulated over a period of nine months. The analysis revealed the following:

- There were not many answers to the questions asked. We hypothesized that the reason for this is that the sales community work exists within tight schedules and they lack the time to search for the information that they may have access to and send a reply. Services professionals have an incentive to prioritize activities that can be billed to an account rather than respond to emails.
- Most requests end with the requester communicating with someone they know and then exchanging information with them. It is critical to find the right person who is involved in a similar situation or can share tacit and explicit knowledge. Therefore, the social network is very important.
- The concepts involved in a request are often complex and abstract. The first email of each thread is more than a couple of sentences long, and we believe this is why typical keyword search solutions are not satisfactory.

The findings were also confirmed by direct interviews that were conducted with the sales community professionals that led to a closer examination of the information requirements of the practitioners involved in business processes. Prior research in search technology [1][2] provides the following additional insights regarding the requirements:

- The business workforce, in carrying out their day-to-day job roles, look for business-related

information in specialized portals and using enterprise search solutions that are based on Internet search tools [1]. However expectations don't measure up and the disappointment contributes largely to the dismal satisfaction reports on search tools for enterprise.

- Business data has security and privacy constraints. Since the current search solutions adopt a conservative approach towards information access, an enterprise user would never know the existence of useful information that will allow him/her to pursue access to it.
- For most business practitioners, the end goal is not just to get access to a single relevant document, but to obtain insightful relationships within the business context that transcend beyond a single document. In many cases, such insightful information leading to the right people within the business is considered a more successful search.
- The contextual presentation of search results is critical to enable successful information discovery and insights. Typically, users want to explore, map, correlate, and analyze the search results so that the answers to their queries make real sense, which is lacking in most of the existing search tools, where the focus is mainly on retrieving a set of "most relevant" document links [2].

Against this background, we introduce a new approach called *business-activity driven search* where the goal is to identify the most relevant and "successful" business activities and get to the information pertaining to those activities along with the people that were involved, rather than just retrieving a document or a chunk of text alone. In general, business activity implies an organized activity that is either all or some part of an enterprise operational process. A sales effort that a service provider might engage in response to an RFP (Request For Proposal) or the development of a particular version of a software product are good examples of a business activity.

2. Business-Activity Driven Search Rationale

To further understand the rationale for business-activity driven search, consider an IT services provider, Universal Services Business (USB), and one of its employees, Jane Smith, who is responsible for selling IT services in the Americas. Jane and her team are working on an opportunity with a banking client in the Financial Services sector. In order to win the business opportunity, the team has to understand the client's requirements and develop an appropriate technical contract (a *deal*) [3] between USB and the

client. In practice, such deals are non-trivial because they typically involve complex assets such as hardware, software, and in some cases, worksites and/or an entire workforce of the client in a department or a site. Consequently, Jane's team has to analyze the business opportunity from multiple viewpoints that require accessing, integrating, and processing different types of information, which includes consulting previous relevant deals to determine such things as:

- Business and technical experts who participated in a deal
- Innovative technologies and business processes that differentiated USB from its competitors
- Most appropriate win strategy for this client
- Pricing models to be utilized
- Customer references

USB stores these documents in several repositories with access control mechanisms that limit access to participants on a need to know basis. Jane knows that the required information about previous deals exists, possibly scattered in multiple places and the experts that worked on them are around somewhere to help her, but she has a realization that her task is not easy to fulfill due to several technical challenges.

First, existing search tools do not connect well with the services processes, i.e., they may not return *actionable* information which Jane can readily use and act upon. Instead, in most cases, Jane is faced with tools and search techniques that deliver results as a set of document links, possibly ordered based on relevancy on what Jane chooses to express with keywords. Thus, it becomes her responsibility to analyze the documents, cull out irrelevant details (including confidential or proprietary information specific to the previous deal) and develop insights by correlating multiple search results.

Second, keyword search is not enough. For instance, Jane may be interested in finding "win strategies" from previous relevant deals. However, since "win strategy" is a services business concept comprising of multiple elements such as the deal pricing model and innovative technical solutions, keyword search on "win strategy", "win plan", etc. will potentially miss many hits and/or provide irrelevant information. In addition, information about complex concepts may be scattered across multiple documents, which require intelligent ways of processing and integrating the data; a feature that keyword search techniques cannot provide.

Third, due to the need to protect client data, access to the raw documents is limited to individuals associated with the deal to which the document pertains. Consequently, with typical search solutions, Jane would get search results from only the deals she

has access to, which might not be helpful at all depending on her need. Instead, knowing the people that worked on relevant deals would give her an opportunity to leverage their experiences and tacit knowledge, but without knowing the business context of the deals she cannot judge the relevancy, and hence has no clue whom to contact. Eventually, Jane may decide to use the email distribution list and upon receiving unsatisfactory answers, she would end up using her personal network of colleagues to seek help. As an experienced sales executive she has a strong social network, but she definitely struggled when she was new to the company.

In light of this scenario, it can be seen that Jane needs a search tool that aligns itself with the business process she executes, lets her search based on the concepts that are important from the viewpoint of her business role, presents results in a granularity that enhances information gain along with the business context of each result in that granularity (facets of a deal), assists her in making the right collaborative connections and presents a set of relevant documents after the interest has been established on a particular deal. Note that, such a search tool will potentially perform better compared to generic search techniques as a result of the following factors:

- Semantic scope of business activities (e.g., sales) are limited
- Concepts related to a business activity are also limited
- Important concepts can be identified from the business practitioners

A focused and narrow search approach that enables retrieval of information only pertaining to specific business activities can be more successful. Such a search tool will typically take inputs such as geography (“americas”), service (“networking”), sector (“financial”), industry (“banking”) etc. and returned deals (most recent on the top) with the business context of each deal exposed, and upon Jane’s interest in the deal, provide a list of business experts who worked on it along with the relevant documents (permitting access privileges).

3. Business-Activity Driven Search

There are two critical elements to the business-activity driven search. First, the user is presented with a set of 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. Identification of important concepts to be presented for a user community is a part of the methodology. Second, as a result of the search query, the search returns a set of the most relevant business

activities first rather than documents or links. A synopsis 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 might limit what a user can see based on their role and access controls on the data.

Since the methodology (cf. Figure 1) is centered on the notion of a business activity that is pursued by different practitioners, a significant element of the approach is to first identify different types of information requirements pertaining to a particular business activity which can be captured by specific *semantic concepts*. For example, earlier we outlined some of the information needs of Jane Smith who is responsible for a particular business activity, i.e. selling IT services. These needs should be determined by interviewing the practitioners or by analyzing data such as an email distribution list. The next step is to build information extractors and document tagging mechanisms referred to as *annotators* [4]. They analyze and retrieve information pertaining to the semantic concepts and as appropriate build semantic indices to enable search.

Annotators contain the logic to identify which documents within the repositories are key to a business activity, and within those documents, what segments should be analyzed for retrieving the required information, e.g. details of a win strategy. Typically, the relationship between a document collection and a business activity would be explicit. For example, a document collection is associated with a particular deal by means of a metadata field that identifies the deal that the document corresponds to. The information thus extracted by the annotators is processed and integrated into a structured knowledge database, which forms the *business activity index*. This database contains information organized by the business activity, including extracted information associated with the business (semantic) concepts, people involved with the activity, and the business context.

In addition, the annotators also add semantic tags or annotations to relevant portions in the document text. The documents together with the annotations result in a *semantic index*. This index serves to connect the business concepts and the documents contributing to them. In a typical semantic search solution, this is the index on which queries are run and hence the result is only the documents that contain those concepts. But in this methodology, the query is issued in terms of business concepts (e.g., networking service) or people,

and is run against the business activity index. This results in retrieval of relevant business activities first (e.g., sales contracts) and the business context of the activities, which includes everything that the business practitioner needs to know to understand the relevance of the activity. At this point the documents themselves do not add any value to the search. Once the relevant and interesting business activities are determined, then the search results are provided at a lower level of granularity (i.e., document links). However, if access to the documents is not allowed for the user, he/she can still leverage information about individuals involved in the activity for a follow-up to pursue access permissions or acquire knowledge directly from the individuals.

The relevancy of the document links is another interesting aspect in this methodology. Since the search proceeds from business context identified by

the practitioner to the documents, the documents that contributed to such business context in conjunction with the original query would be the most relevant. For instance, if a practitioner looks for *win strategy* and selects a deal, then the document links from that deal that contributed to the *win strategy* would be the relevant documents. The same could have been obtained by a semantic search; but the knowledge that the practitioner develops on the way to finally getting the documents is the key difference.

Another aspect of the methodology is to enable keyword search, sometimes called *general search*. This is motivated by the cases where practitioners are looking for something unique and hence will not be able to use concept search by itself. Such cases are handled by enabling search for keywords and phrases that runs these queries against the inverted index, which is a part of the semantic index, to obtain ranked

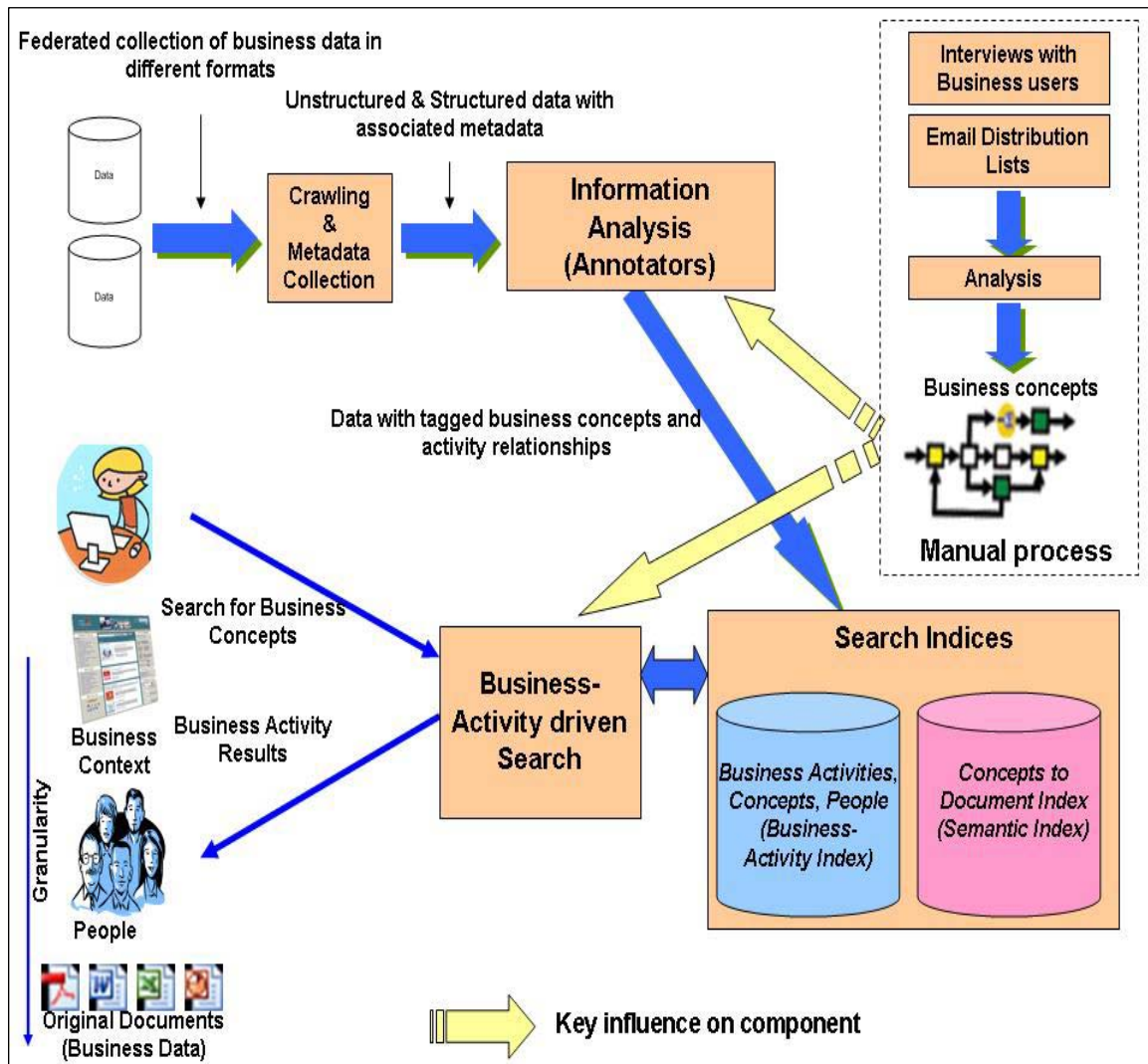


Figure 1 Business-activity driven search methodology

document links and displays the search results with the business activities that the documents belong to. Search for keywords could also be used in conjunction with the concept search to first narrow down the most relevant activities according to keyword search, and then apply the concept search criteria to the previous results. The keyword would be shorter in this combined “general search” than in typical keyword searches, because most of the key concepts that contribute to confusion with large sets of keywords are already removed by having well-defined concepts. Hence the precision of the results can be expected to be better.

Several research efforts have been proposed for information management in business that apply various techniques including analysis of unstructured information from multiple repositories [5], combination of structured and unstructured information analysis [6][7], ontology-based text analytics that address semantic heterogeneity and disambiguation [8][9], and machine learning approaches to evolve semantics and knowledge discovery from text [10][11]. However, none of the existing techniques in isolation address the bigger problem which is the ability to exploit the results of text analytics, mining and semantics for the benefit of a business process, and in the services domain.

There is also related work in the area of user driven search which addresses the granularity of search and the benefits of focusing at sub-document levels [12]. Our approach increases the granularity to business activities rather than going down to sub-documents. However, we use sub-document level extracts to deliver insights that provide the business context for the particular business activity.

Faceted search is often employed to provide the ability to drill down into search results based on certain independent criteria. For example, Knabe and Tunkelang discuss a wide variety of query refinement techniques from a user experience perspective including the faceted search [13]. Our approach can be thought of as a non-trivial use of faceted search leveraging information extraction. It is fundamentally different from all prior work in the way it combines semantic search with services processes and social networking. In addition, the use of business-activity as a context is particularly novel.

4. Prototype Implementation

A prototype enterprise search solution based on the methodology suggests that an implementation of the above described methodology requires a flexible enterprise search platform that allows text analysis capabilities in its data processing pipeline and plug-in capabilities to customize some aspects of the pipeline.

It also requires a sophisticated annotator development tool. Since the quality of the overall system critically depends on the quality of annotators as well as the efficiency of their development and validation process, the annotator development tool is an important technology required for this methodology.

The significance of the annotators in this search solution can be seen through their role in the solution in some detail. Annotators process the text from the documents and associated meta-data to produce additional meta-data (or annotations). These annotations are further analyzed (by additional annotators) to extract relevant information for a particular business activity (e.g. contact phone numbers of key participants in the activity), which could potentially exist in multiple documents. Some annotations specifically provide the context to the business content and are used only for establishing relevance at the business activity level. For example, the “geography” pertaining to a deal contributes more information gain at business activity level than at the document level. Some other annotations are used at the document search level. For example, the “technical solutions” in a deal are not only important to provide the context but also to help narrow down to the documents that contain them, and hence are part of the semantic as well as business-activity indices.

In most cases, development of annotators is an on-going process due to changes in information requirements of business practitioners as well as for improvement of the quality of information extracted or tagged. The technique used to build and deploy these annotators depends on several factors such as the consistency of the documents created, stability of semantic concepts and corresponding heuristics, existence of domain specific taxonomies and/or ontologies, availability of subject matter expertise, and the existence of business rules and conventions to enable simpler extraction and tagging. For instance, annotating named entities is well researched and this could be leveraged in the context of a semi-structured business document with known conventions. If the title of a particular type of PowerPoint presentation has customer name on it because of naming conventions, then applying named-entity detection on the title would yield powerful results. For more abstract and complex concepts such as win strategy, machine learning techniques would be best suited if enough annotated training data is available.

For our prototype, we leveraged IBM Unstructured Information Management Architecture (UIMA) [5] platform for building a range of annotators based on various techniques including regular expressions, domain heuristics based, semi-structured information analysis, ontologies, and text classifiers. We also used IBM WebSphere Information Integrator OmniFind [4]

for its robust, scalable and flexible enterprise search platform capabilities. This includes the pipeline of crawling, parsing, indexing, and search runtime.

OmniFind also utilizes UIMA components to perform some standard analysis of the text including language identification and lexical analysis.



Figure 2 Screenshot of the prototype showing the search query input panel and sample results.

In the prototype, the users interact with the search system via a user interface that exposes semantic concepts (cf. Figure 2) relevant to their role. A user query (inputted from the interface) is converted into a set of SQL queries on the business activity index first, and later SI-API (Search and Index API) [4] queries to the semantic index. The SQL queries extract the first level results, which are business activities relevant to the user query and for each business activity it retrieves the business context of the activity (from the database itself). Furthermore, the relevant documents within the scope of that activity are extracted from the semantic index using SI-API queries. Hence, when a business activity is selected by the user, the document links under that will be the key documents that contributed to the relevance of that activity.

Referring back to the earlier example, if Jane were to use the prototype for her information needs, she would select “Networking” (from a pre-defined taxonomy) as one of the “Towers” (or services) in the deals she is interested in, “Sector” would be “Financial Services,” Industry would be “Banking” and “Geography” would be “Americas”. Once these inputs are provided, the system will identify those deals that satisfy the query (cf. Figure 2). Assuming that multiple deals were retrieved in return to Jane’s query, she can first view the business context (synopsis) of the deals which contains various information related to the particular deal (e.g. win strategy, list of key contact persons), and if interested, retrieve the original documents pertaining to that deal (permitting access privileges). Thus, the goal is to present the user with a unified view of the information pertaining to a particular deal that will allow the user to correlate and make an informed decision about whether s/he wants to delve deeper and look for more information either by retrieving the original documents (from which this information was extracted) or by networking with the people associated with the deal. If she chooses to communicate with someone that worked on a deal, the prototype enables integrated access to the corporate directory. The search methodology, therefore, should be a part of and fully integrated with the existing information access and collaborative tools.

5. Challenges

There are several research and technical challenges introduced by the above methodology which proposes an innovative combination of information extraction (IE) and information retrieval (IR) to support information discovery within a business context (specific business domain). A significant challenge is to develop evaluation metrics that will enable comparison of this methodology with existing

information retrieval or extraction systems. Primarily, during such a comparative evaluation, the precision/recall for these systems must be calculated at the business activity level instead of a document level, because there is significant information gain by reaching the relevant business activity. Also the queries as well as the results are different, hence the comparison should also be in terms of number of query runs and document scans necessary in other IR systems to achieve the desired information gain for a typical business activity related search.

Preliminary findings from a comparison of this methodology with keyword/key-phrase search techniques shows a significant amount of noise reduction, i.e., the user had to scan through a significantly fewer number of documents to achieve the desired information gain. A broader user evaluation is needed to obtain feedback on how well the concept selection represents the typical information needs, how effective the retrieval is in terms of reducing the amount of time needed to get to the “right” information needed to perform the business task, how effective the business context of an activity is in terms of providing the “right” information without further deep dive searches, how it has helped build the right social network to accelerate the business activities, and how it compares to keyword search for a selected set of typical business activity based scenarios in terms of precision/recall. The overall feedback determines the effectiveness of the search methodology instead of just the standard precision-recall curves. We believe this will establish a promising direction for exciting future research and enterprise search.

Furthermore, there are several aspects of the methodology that lead to interesting research problems which includes ranking of the business activity results based on business-identified goodness measures (e.g., ideal activities resulting in wins having higher rank), techniques to improve specific annotator accuracy to improve overall system performance, techniques to build probabilistic measures of individual and composite annotator accuracy to influence the retrieval and ranking, approaches for improving system performance based on user evaluation and feedback at business activity granularity (business-activity driven recommender systems), and extracting business context from the document-level semantic annotations (building a collective view of business information). Above all, the process of identifying important semantic concepts for a particular practitioner role in a business process is perhaps one of the most challenging issues. Automated means of identifying, even an initial set of these concepts, and would prove to be very useful in this methodology.

6. Conclusion

In this paper, we introduced business-activity driven search—a novel approach for search and information management within enterprises. The approach is based on two key observations made about information seeking behavior in enterprises: (i) a limited set of semantic concepts represent the critical information needs of business practitioners based on their business role, and (ii) enterprise information management is centered around organized business activities (e.g., sales) and practitioners of those activities. Consequently, our approach consists of identifying key concepts that are representative of a particular activity and developing text analytics (called annotators) for extraction of information about the activities (i.e., the business context) and identifying key persons involved in those activities. A unique feature of this business-driven search methodology is to allow its users to first retrieve the most relevant business activity (based on business concept-based search criteria) along with the business context and people involved in the activity, and then access the relevant documents from that activity permitting appropriate privileges. Note that, such an approach is in stark contrast to the traditional enterprise search techniques where the users are often confronted with a set of document links as search results, instead of a more focused and connected view of information that are most relevant to their business role. Indeed, this methodology may not be suitable to handle arbitrary Web (or Internet) search queries. However, a proof-of-concept and preliminary evaluations have shown the promise of such an approach in a restricted enterprise user community that has specific information requirements. This demonstrates that the proposed approach goes beyond a simple listing of text results (e.g., document links), and enables users to explore commonalities between groups of search results which helps in building a connected picture of those results from seemingly unconnected data.

Acknowledgment

The authors would also like to acknowledge the system design, integration and development efforts of Norbert Vogl, Jim Rubas and Heather Smith as well as thoughtful discussions in the process of prototype development. The authors are thankful to Howard Sachar for introducing us to this information management problem and his continued guidance in making this research useful for practitioners.

References

- [1] R. Mukherjee and J. Mao. *Enterprise Search: Tough Stuff*. ACM Queue, 2(2):36–46, 2004.
- [2] R. Lewis *Zen and the Art of Enterprise Search*. CIO Magazine, September 2006. URL: <http://www.cio.com/weighin/column.html?CID=24882>
- [3] H. Chesbrough and J. Spohrer *A Research Manifesto for Services Science*. Communications of ACM, 49(7):35–40, 2006
- [4] N. Alur, T. Brown, C. Delgado et al. *WebSphere Information Integrator OmniFind Edition: Fast Track Implementation*. IBM Redbooks, 2006.
- [5] 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.
- [6] P. Dmitriev, N. Eiron, M. Fontoura, and E. Shekita. *Using Annotations in Enterprise Search*. In 15th World Wide Web Conference, pages 811–817. ACM Press, 2006.
- [7] 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.
- [8] 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.
- [9] 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.
- [10] 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.
- [11] 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.
- [12] X. Li, T-H Phang, M. Hu, and B. Liu, Using Micro Information Units for Internet Search, Proc. of the 11th ACM International Conference on Information and Knowledge Management, pages 586–573, Nov 2002.
- [13] F. Knabe and D. Tunkelang, *Enterprise Information Access and the User Experience*, IEEE IT Professional magazine, pages 21–28, January-February, 2007.