RI 13004, September 2013

IBM Research Report

A Case Based Approach To Serve Information Needs In Knowledge Intensive Processes

Debdoot Mukherjee
IBM Research-India, New Delhi, India

Jeanette Blomberg
IBM Almaden Research Center, San Jose, CA, USA

Rama Akkiraju
IBM Almaden Research Center, San Jose, CA, USA

Dinesh Raghu
IBM Research-India, Bangalore, Karnataka, India

Monika Gupta
IBM Research-India, New Delhi, India

Sugata Ghosal
IBM Research-India, New Delhi, India

Mu Qiao
IBM Almaden Research Center, San Jose, CA, USA

Taiga Nakamura
IBM Almaden Research Center, San Jose, CA, USA

IBM Research Division: Almaden - Austin - Beijing - Delhi - Haifa - T.J. Watson - Tokyo - Zurich

Abstract. Case workers who are involved in knowledge intensive business processes have critical information needs. When dealing with a case, they often need to check how similar case(s) were handled and what best practices, methods and tools proved useful. Since popular Case Management tools do not have mature knowledge management capabilities, case workers must look for necessary information using various enterprise search systems and information portals. However, state of the art enterprise information systems fail to deliver relevant and contextual content to case workers. This sentiment is confirmed by our primary research conducted by interviewing knowledge workers engaged in the Opportunity-To-Order process at a large IT services company. In this paper, we present our Solution Information Management (SIM) system developed to assist case workers by retrieving and offering targeted and contextual content recommendations to them. In particular, we present a novel method for intelligently weighing different fields in a case when they are used as context to derive recommendations. Experimental results indicate that our approach can yield recommendations that are approximately 15% more precise than those obtained through a baseline approach where the fields in the context have equal weights. SIM is being actively used by case workers in a large IT services company.

1 Introduction

Case Management [30] has emerged as the discipline for supporting flexible and knowledge intensive business processes, which may require significant human judgment and decision making. Unlike traditional Business Process Management (BPM), which has focused on automating process workflows, Case Management is aimed at equipping *knowledge workers*[1] efficiently steer processes toward completion. Since, knowledge workers add significant economic value to an enterprise and their contributions are especially critical in growing the services economy, the demand for Case Management tools has been growing [3]—especially in domains such as customer relationship management, IT service management, healthcare, legal, insurance and citizen services. When knowledge workers begin to work with a case they often ask—*Did we handle such a case before? If so, how? What best practices are available to solve similar cases?* To get answers to such questions, they often search in enterprise repositories. However, knowledge workers are frustrated with the inability of the available knowledge management tools in finding the information they need, when they need it, due to the poor state of the art of enterprise search. Studies report that they may spend 15% to 35% of their time searching for information and are successful in finding relevant information less than 50% of the time [13, 11]. In practice, what works better is reaching out to subject matter experts in the organization through informal networks. But, identifying the right person often requires numerous phone calls and email exchanges, which takes up precious, productive time of knowledge workers. Clearly, developing technologies for effectively aggregating and disseminating case knowledge is a strong business imperative for next generation Case Management [20, 27, 29] and Social BPM [28] products.

State of the art Case Management tools enable knowledge workers to arrive at faster and better decisions by presenting a holistic view of a case. They also recommend similar past cases based on business rules conditioned on properties of the current case or the workflow step. Such features have found tremendous applicability in processes such as claims processing, credit approval and dispute management. For processes with more complex knowledge needs, which require creative thinking or research, case workers are directed to enterprise information systems (*e.g.,* enterprise search, domain-specific content management or document management systems). However, we find that such Information Retrieval (IR) systems are not at all effective in addressing the information needs of case workers. Our study of knowledge workers at a large IT services organization reveals that a multitude of technical and organizational challenges currently make

[1] The term, *knowledge worker*, was first coined by Peter Drucker to denote those who develop or apply knowledge in the workplace. [23] discusses different roles of knowledge workers.

it extremely difficult for case workers to find the information necessary for their daily work. One key issue is that the IR systems are not closely integrated with the Case Management frameworks; hence the search results are only as good as the keywords entered by the users. Therefore, a new and focused approach to knowledge management is required that is aware of prior cases and their context to provide effective assistance to knowledge workers. This paper describes how information retrieval guided by the context of the case-at-hand and the semantics of the case domain generates useful content recommendations for the knowledge workers.

In this paper, we present a novel approach to knowledge management leveraging Case Management principles and discuss its application to the *Opportunity-To-Order* process at an IT services company. Central to the activities of *Opportunity-To-Order* teams, is the collaborative development of a solution in response to a Request for Proposal (RfP) issued for an IT services project. Different roles are entrusted with different responsibilities. For example, a *Solution Architect* creates a solution to fulfill the client's requirements based on the service provider's delivery capabilities and offerings, a *Risk Manager* assesses the risks inherent in the assumptions made during solution design, a *Pricer* prices the proposed solution, and a *Client Sales Executive* drafts the response to the questions raised in the RfP. These activities often require researching what has or hasn't worked in "similar" cases and studying relevant best practices and reference materials. At the outset of this project, we interviewed knowledge workers engaged in the *Opportunity-To-Order* process at a large IT services organization to learn about their information seeking practices. They expressed deep dissatisfaction about the state of the art for information retrieval and were extremely keen on seeing breakthroughs in information delivery methods. Here are some of the notable themes in their wish-list:

– Having a single location for search instead of dozens of content portals.
– Getting precise results that are contextually relevant to the current case.
– Access to crisp, summarized information that leads to targeted insights.

To address the above noted requirements, we developed a knowledge management solution called Solution Information Management (SIM). SIM mines contextual and targeted information by searching a federated set of repositories. The repositories store solution design documents created during past opportunities as well as best-practice reference materials about offerings, delivery capabilities, lessons learned and engagement process. A *case* uniquely identifies an IT service deal that was pursued in the past and for each case, we catalog information related to the deal in a *case model*. Different fields of the case model contain metadata about the deal such as the client name, geographies that were involved in delivering the services, the type of services to be delivered to the client in that deal, the industry of the client, the size of the deal as measured by the total contract value, the complexity of the deal as measured by the number of services involved in the deals etc. We apply an array of information extractors to resolve the semantics of content in the unstructured documentation created for the deal. For example, we extract information on the win themes and value proposition used in the deal, the architecture of the solution, and the Service Level Agreement (SLA) that was in place. All of these are also added as fields in the case model for the deal. Then, the integrated case models are indexed for search purposes. Such a richly fielded index allows the definition of targeted semantic queries and not just keyword search on full text of documents. Suppose, a case worker is looking for existing prior assets or lessons learned on "low cost data center consolidation solutions in financial service industry in Western Europe". In SIM, one can create a set of query clauses on different fields to address such a requirement—*Geography* : "Western Europe", *Offering* : "data center consolidation", *Win-Theme* : "Low Cost", *Industry* : "Financial Services". The results obtained from a such a query are much more precise than what a keyword search would yield. Further, the SIM system can generate content recommendations for the information needs in different process steps based on the already known fields in the case or the *context* of the case. An interesting question that arises is how to weigh the affect of different fields in the context. In the above example, suppose the case worker is now interested in recommendations for potential risks underlying the solution. How do we weigh the four query clauses as we look to retrieve cases with *Risks* that can be of interest? Do we weigh the clause on *Offering* more than the others or is a match of the *Industry* more important to fetch relevant *Risks*? Resolving an appropriate weighting of the different query clauses is crucial in order to maximize the relevance of recommendations. It is a complex problem since there are hundreds of fields in industrial case repositories and manually specifying the weights of different fields as they are leveraged in deriving recommendations for other fields is infeasible. We propose an automated approach, named *Correspondence Analysis*, which infers how one case field can influence recommendations for another case field. For each pair of case fields (X, Y), we analyze pairs of cases in the case corpus to understand how the similarity in field X can correspond to similarity in field Y. This correspondence dictates the weight of the query clause formulated from field X in generating recommendations for field Y. In the running example, we determine how often do we observe similar *Risks* across pairs of cases that have the same *Offering*; this helps define the weight for the query clause with *Offering*. We conduct experiments where we assess the relevance of recommendations on two different case fields, obtained from a corpus of 715 cases. The relevance of recommendations obtained through our approach is significantly better than that from a baseline approach which assumes equal weights for all fields in context. Improvement observed in standard IR metrics like *Precision@K* and *nDCG* is as high as 15%.

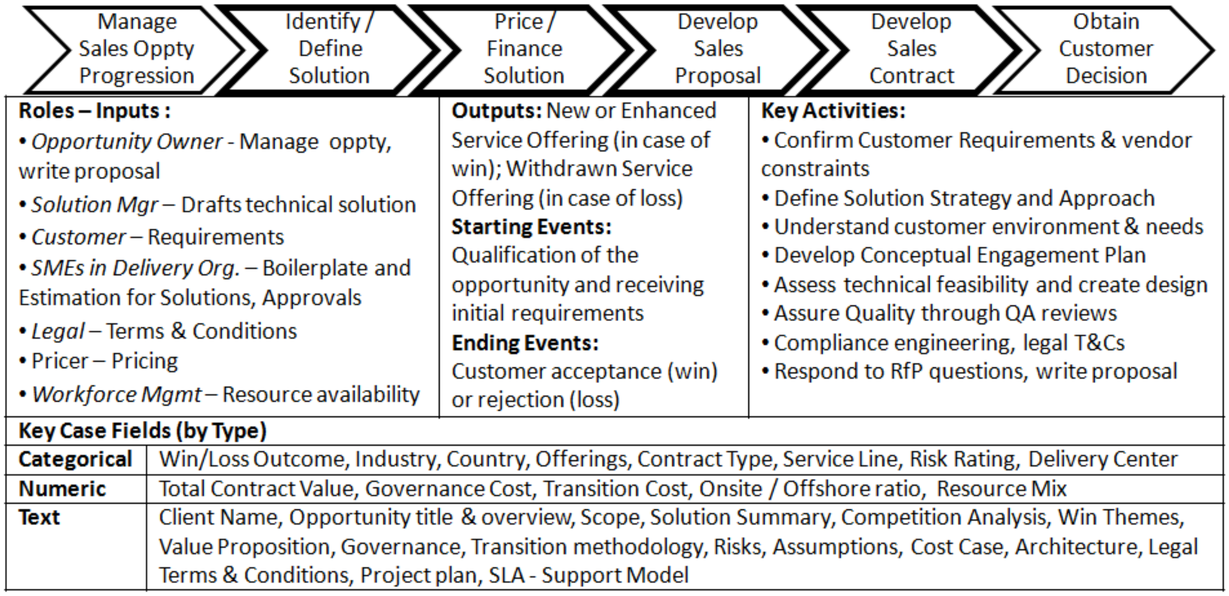
Our contributions include:

- A survey of information needs of case workers who are involved in the knowledge intensive process of Opportunity-To-Order.
- A description of an end-to-end knowledge management system that delivers contextual and targeted information to case workers as they need it.
- A novel approach for learning a preferential weighting scheme for fields in contextual queries through an analysis of the case corpus.
- An experimental evaluation of the efficacy of our contextual query weighting technique in improving relevance of recommendations.

The remainder of the paper is organized as follows. Section 2 presents our study of knowledge workers and describes the domain in which they work. Section 3 describes the system architecture and details of our knowledge management solution. In Section 4, we present the results of our experiments to evaluate the efficacy of the context weighting technique, followed by a brief overview of related work in Section 5. Finally, we conclude with a discussion on future work.

2 Opportunity-To-Order Domain & Knowledge Worker Study

To better understand the information needs of knowledge workers, we conducted primary research within a large IT services organization, specifically focusing on the Opportunity-to-Order (O2O) process. In this section, we introduce the readers to this process, then outline the current support systems within the organization and finally report the findings of our primary research.



Roles – Inputs :	**Outputs:** New or Enhanced Service Offering (in case of win); Withdrawn Service Offering (in case of loss)	**Key Activities:**
• *Opportunity Owner* - Manage oppty, write proposal • *Solution Mgr* – Drafts technical solution • *Customer* – Requirements • *SMEs in Delivery Org.* – Boilerplate and Estimation for Solutions, Approvals • *Legal* – Terms & Conditions • Pricer – Pricing • *Workforce Mgmt* – Resource availability	**Starting Events:** Qualification of the opportunity and receiving initial requirements **Ending Events:** Customer acceptance (win) or rejection (loss)	• Confirm Customer Requirements & vendor constraints • Define Solution Strategy and Approach • Understand customer environment & needs • Develop Conceptual Engagement Plan • Assess technical feasibility and create design • Assure Quality through QA reviews • Compliance engineering, legal T&Cs • Respond to RfP questions, write proposal

Key Case Fields (by Type)	
Categorical	Win/Loss Outcome, Industry, Country, Offerings, Contract Type, Service Line, Risk Rating, Delivery Center
Numeric	Total Contract Value, Governance Cost, Transition Cost, Onsite / Offshore ratio, Resource Mix
Text	Client Name, Opportunity title & overview, Scope, Solution Summary, Competition Analysis, Win Themes, Value Proposition, Governance, Transition methodology, Risks, Assumptions, Cost Case, Architecture, Legal Terms & Conditions, Project plan, SLA - Support Model

Fig. 1. Opportunity To Order: Inputs, Outputs, Events, Activities, Case Attributes

Background & Objectives: During the *Opportunity-To-Order (O2O)* process, knowledge workers in different roles collaboratively create a solution that is in response to a request for proposal issued by a client. With response time frames getting shorter, IT services companies are under tremendous pressure to quickly turnaround proposals. To accommodate such tight deadlines, the productivity of the O2O sales work-force needs significant improvements. Analysts [10] suggest that if social technologies can enable knowledge workers to effectively search and analyze information and quickly connect them with experts, then their productivity can improve by as much as 20-25%.There is also a strong incentive to leverage the collective intelligence of an organization, including lessons learned from past engagements to improve solution quality. A comprehensive knowledge management system for example can enable solution architects to support their design assumptions with relevant and timely data points, lessening their reliance on what they have done before by providing access to the experiences and successes of others. Motivated by this need, we set out to understand the information needs of pre-sales and solution design communities at a large IT services organization.

State of Process: The *Opportunity-To-Order (O2O)* process encompasses the sales lifecycle of a services engagement. Sales activities start when the client issues a Request for Proposal (RfP) and continue until a service delivery contract has been signed by the client and the winning vendor. Figure 1 presents a high-level overview of the process as enacted at the large IT services organization where we conducted our research. It describes the key inputs coming from different roles and the main activities that are performed. Although this process appears to be linear, in practice a great deal of iteration and collaboration characterizes the activities of those responsible for the various stages of the process [15]. A wide range of information gets documented in each *O2O* case. Figure 1 lists the important attributes in such a case. Whilst some of them are well-defined categorical fields, most of them are entered as free form text. At the time of this study, our subjects were using Lotus Notes databases to store the structured attributes (categorical and numeric) in cases and Microsoft Office tools (*e.g.,* Powerpoint, Excel, Word) to author dense, richly formatted documents containing free form text and images[2]. Creating content for most of the case attributes requires significant research on past cases and the company's offerings and delivery capabilities. Ineffective search features in the case tools results in case workers having to rely on their personal networks to gather information about similar cases. While personal networks can be a great resource they provide access to only a small fraction of the people in the enterprise who may have valuable information. Furthermore in settings where there is rapid growth or high turnover people don't have the time to develop rich networks. Finding the right people may involve numerous phone calls and emails with mixed results if the experts have left the organization or are unable to help given their time constraints. Complicating the task of finding relevant information on offerings and capabilities is the fact that this information is scattered across numerous portals and repositories making it difficult to locate simply by using current enterprise search capabilities. However without this critical information, case workers must make decisions that may not be optimal for the current business or technical context. To better understand the issues and the information needs of case workers in O2O process at the IT services organization that we worked with, we conducted a set of interviews.

Study Method: We conducted open-ended interviews with 33 knowledge workers engaged in O2O activities. In addition, we conducted user testing on an early version of the SIM prototype with 4 technical solution design managers. Participants in the study represented different levels of expertise and different levels of seniority within the organization. The focus of the open-ended interviews was on the work practices of the interviewees, particularly in relation to finding context relevant information required to advance the O2O case. Specifically, we asked questions about participants' roles and responsibilities, their information needs broadly defined, strategies for locating, accessing, and managing relevant information, and their approach to creating and sharing information with others. The open-ended style of the interviews meant that while an interview protocol guided the discussion, we encouraged interviewees to report on their everyday experiences to provide us with examples of the strategies and approaches that they used to manage their information needs. Interviews, lasting approximately 60 minutes, were audio recorded and detailed notes were taken during the interviews. The notes and selected audio recordings were reviewed as input to the analysis which consisted of identifying recurring topics, grouping the topics into themes, and associating specific evidence (*e.g.,* verbatim or paraphrased quotes) and supporting materials with each theme to underscore the importance and meaning of the theme.

Study Findings: Our findings are grouped into four issue areas, each of which has guided the design of the SIM tool. We discuss these areas highlighting the perspectives of those interviewed.

Awareness and Access: Two related issues were identified as limiting the ability of O2O practitioners from taking advantage of the collective knowledge of the organization. First, interviewees reported that practitioners were sometimes unaware that certain information existed. This resulted in reliance on potentially outdated or incomplete information. Furthermore, for some the way the information was presented limited their ability to understand its relevancy to the particular task at hand. Second, even when interviewees suspected that there might be relevant information, they were not able to easily access the information. Information was "walled-off" in repositories created specifically for geographically localized or specialized teams with access formally or informally restricted to members of these teams. Special permission had to be granted to access some repositories (and access control policies differed by geography

[2] One practice had initiated efforts to switch to an Adaptive Case Management tool.

and expertise area). If direct access was not possible, documents had to be forwarded by someone with direct access.

Networking and Sharing: Not surprisingly interviewees reported that they relied on their personal networks to find information. These networks however had limitations based on a person's tenure or particular role in the company and on their geographic location. For some, their networks were confined geographically, restricting the global flow of information. This also meant that there was differential access to experts given that in some capability areas the concentration of expertise varied by geography. Sharing information through personal networks allows practitioners to access information that is known or managed by people in their network. It is also the case that information is pushed out to practitioners, usually by division-wide groups who support the O2O practitioners. While a push strategy can create awareness about the existence of information, there is a risk that the information may not be relevant to practitioners at the time it is made available. A balance is required between pushing information out to practitioners and making it easy for them to pull information from information sources.

Quality of Information: Once information has been identified and access granted, practitioners have to determine if the information is to be trusted. Information ages at different rates which can diminish its value overtime and in some cases render it detrimental to the success of the O2O case. For example, information such as costing data may become out of date quickly as new technology introduction tends to diminish the value of older technologies. On the other hand, information about the acquisition history of a particular client may be relevant for years. Practitioners develop strategies to determine if they have the most recent, up-to-date information such as looking at the dates on which documents were created or noting clearly outdated product names rendering the rest of the information in the document suspect. Beyond simply attending to the age of a document, interviewees cautioned that there were risks associated with relying too heavily on information about past successes. Equally important was learning about new innovations or strategies not yet tried that might address the requirements of any particular O2O case.

Work climate: O2O practitioners stressed that they and others worked in a very fast paced work environment where there were serious constraints on the time people had to search for and adapt information for their immediate need. Whatever solution was developed to help with their information requirements, it would need to be easy to use, available 24X7 and allow for the rapid identification of relevant information. Adding to the challenges of the work environment was rapid growth and relatively high turnover in some geographies resulting in a high concentration of new employees who had to come up to speed quickly and who relied on easy access to information resources. The solution needed to address the particular demands of this population along with their more seasoned colleagues.

3 System Overview: Solution Information Management

In this section, we describe Solution Information Management (SIM), a knowledge management system that we developed to address the information needs of case workers in the Opportunity-To-Order process at a large IT Services organization. The SIM tool assembles content from a variety of data sources, indexes the information after converting it into a semantic format, and then delivers relevant information to the case worker depending upon the context of a case. Figure 2 shows the different stages in the knowledge engineering pipeline in SIM—*Crawl & Parse*, *Annotate*, *Query*, *Index*, *Search*, and *Visualize*. Now, we briefly describe our implementations for these stages and discuss the key design issues.

Crawl & Parse: As discussed in Section 2, knowledge workers in the Opportunity-to-Order process need to visit a large number of repositories and information portals in order to satisfy their information needs. For each relevant data source, we configure a crawler in SIM that periodically downloads contents of the repository. Most Enterprise Search products have crawlers for popular database technologies (*e.g.,* Notes, Sharepoint, FileNet). For crawling the Lotus Notes based case repositories, SIM uses the Notes crawler implementation from IBM Content Analytics [4]. In addition, SIM also crawls different web-based wikis and portals by using focused crawling techniques [9].

The crawlers output files in their native, binary formats, *e.g.,* .pdf, .ppt, .xls, .doc. The next step is to *parse* the text from such files. We employ the Apache Tika [2] parser to extract plain text from the files. In order to feed content into some of our annotators we require the parsed text to preserve formatting and style information—we make use of COM APIs in Microsoft Office for such advanced parsing needs. Also, we export pages and slides as images; these images show up alongside search results to enable a preview feature (See discussion on *Visualize*).

Annotate: The *Annotate* stage creates semi-structured case models with information extracted from multiple, unstructured text documents associated with historical cases. For instance, O2O case workers create dense solution design documents in Microsoft Powerpoint that describe various categories of information related to solution (*e.g., Win Themes, Value Proposition, Solution Architecture, Service Level Agreements, Risks, Assumptions*). The *Segmenter* module in SIM's *Annotate* stage can convert such formatted documents in proprietary formats into semi-structured (XML) case models, where the texts in the different information categories are captured as case fields. It takes as input the formatted text parsed from the documents. It distinguishes the headings in the documents from any other text based on their special formatting or font-styles. Next, it feeds the words in an inferred heading to a trained text classifier model which predicts the case semantic implied by that heading. A text classifier is trained beforehand by manually labeling a sample of headings by their case semantics. Once we determine the case semantic for a heading, we extract the text from the region following the heading into an appropriate case field. We aggregate case fields extracted from all documents created for an opportunity in a single case model. Also, we map the structured fields available in the case repositories to our case models.

Segmenting an unstructured document and subsequently indexing it as a semi-structured case model with a set of defined fields is crucial for a number of reasons:

1. Queries can yield more relevant results when they are appropriately targeted to a select set of fields as compared to the scenario where queries are processed over whole documents [12, 16, 25].
2. Users can selectively browse (or preview images for) the case fields that they are interested in.
3. The fundamental task of computing the similarity between two pieces of text, which comes up in different steps of our pipeline, may be performed more accurately if the texts being compared are short in length and have common semantics. Imagine, judging the similarity of two long, heterogeneous documents versus that of two short paragraphs—the latter is often much simpler.

We append two additional attributes, namely *Quality Score* and *Summary*, to each case field in our case models which are derived from unstructured text. We obtain these from the modules described below: *Quality Scorer:* Suppose, a case worker engaged in a case, X, is seeking recommendations for the case field, α, based on the context specified by other fields in X. Now, say we find a case Y where the values of all the case fields closely match those in the query created from case X. However, if the field α in Y is sparse or only has boiler-plate text, then it does not make for a good recommendation despite a high match in context. Thus, it is important to gauge the amount of content in any field while we generate

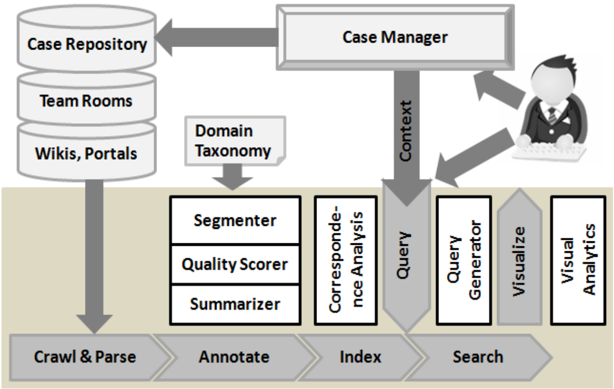


Fig. 2. Knowledge Engineering Pipeline in SIM

```
function CORRESPONDENCE-ANALYSIS
 Input: Case Corpus, $C$; Set of Fields, $F$
 Output: $Corr$ - $|F| \times |F|$ matrix
    Initialize $\Sigma$ - $|C|^2 \times |F|$ observation matrix
    for all $(c_i, c_j) \in C \times C, i \neq j$ do
        Initialize observation vector, $\sigma$
        for all Field $f_k \in F$ do
            $\sigma_k \leftarrow Sim(c_i.f_k, c_j.f_k)$
            Add $\sigma$ to $\Sigma$
        end for
    end for
    for all Field $f_i \in F$ do
        Build a regression model, $M$, from $\Sigma$ to model column $i$ using other columns as features.
        for all Field $f_j \in F$ do
            $Corr(j, i) \leftarrow$ Coeff. of feature $j$ in $M$
        end for
    end for
end function
```

Fig. 3. Correspondence Analysis

recommendations. Our *Quality Scorer* module assigns a score to a text field in a case by assessing the amount of information present in it relative to that present in the same field in other cases in the corpus. For each word in a case field, we calculate its *Inverse Document Frequency (IDF)* as the frequency of the word in that field relative to its frequency in the same field in other cases. *Quality Score* is computed as the average IDF for all words (except stop words) and is boosted by a factor if a field contains attachments, images or domain keywords.

Summarizer: Search results in SIM are accompanied by *summaries* to help users do a quick evaluation of different case fields. A summary for a case field is obtained by choosing a small number of sentences that may convey the maximum information. We select top sentences ranked on the basis of their average IDFs. Thereafter, we use Maximum Marginal Relevance [8] to remove redundancy from the chosen sentences and introduce diversity in the summary.

Index: The case models obtained in the *Annotate* stage are imported into a full text index of a search engine. SIM uses Apache Solr [1] as the foundation for indexing and search. Note that the three consecutive stages, *Crawl & Parse*, *Annotate* and *Index* happen in an offline manner at scheduled intervals. Next, we describe another offline step, the results of which are utilized by the online *Query* module.

Correspondence Analysis[3]*:* As we derive content recommendations for a certain field of a case (henceforth referred to as the *target* field), it is important to understand what other fields in the case can serve as *context*. For instance, as we seek recommendations on the case field, *Risks*, does it make sense to search for *Risks* in cases from the same *Industry* or cases with the same *Solution Offering*? If both *Industry* and *Solution Offering* appear to lend *context* to *Risks*, then how does one weigh the influence of each of these fields? Turns out that even domain experts are unable to conclusively answer such questions. Again, assigning equal weights to the affect of each field in the context does not appear to be ideal (See Section 4). Moreover, since there could be hundreds of fields in case repositories, manually defining weighting schemes may not be feasible. Here, we describe *Correspondence Analysis*, an automated approach that analyzes the case corpus to infer how similarities in different case fields correlate with each other. Also, we discuss how the correspondence output can be used for defining preferential weighting for fields in contextual queries.

[3] Not to be confused with the multi-variate statistical technique with the same name that summarizes categorical data in a two dimensional graphical format

For each pair of case fields, say α and β, we define *Correspondence*, $Corr(\alpha, \beta)$, as the degree to which similarity in α corresponds to similarity in β across pairs of cases. A high value for $Corr(\alpha, \beta)$ suggests that α is a good candidate to serve as *context* for β because if we are able to retrieve cases with similar α, then it is likely that the contents of β in those cases may recur in the current case. For the above example, if our analysis of the case corpus shows that cases with the same *Offering* often exhibit similar *Risks*, then a case worker would be interested to find *Risks* in past cases that have the current *Offering*. Thus, it may be worthwhile to assign a high weight for the query clause with *Offering*. Now, when deriving recommendations for a target field, it is important to have the "right" relative weighting for all other fields in the context. We use multiple linear regression as a tool to determine the relative impact that fields in the context can have upon the target field.

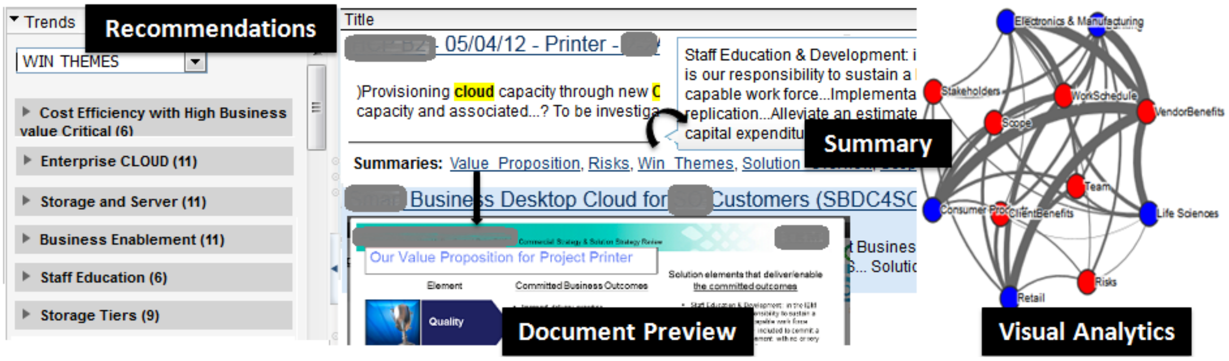
Figure 3 discusses the algorithm for correspondence analysis. We sample pairs of cases from the case corpus to observe how the different fields are similar across the case pairs. The similarity function, *Sim*, depends on the type of field. We use boolean similarity for categorical fields, cosine similarity[4] for text fields and inverse of euclidean distance[5] for numeric fields. Next, in order to assess how similarity of a target field may be influenced by similarities in other fields, we regress the observations from all other fields against the corresponding observations for the target field. The coefficients obtained from a linear regression model can be indicative of how similarities in different fields influence similarity in the target field. Next, we describe the *Query Generation* module that uses these correspondence values to create an appropriately weighted contextual query.

Query & Search: SIM generates specialized queries that are processed on top of the Apache Solr index to derive real time recommendations for case workers. A template for a SIM query is an OR-ed construct of several clauses that are weighed depending upon the target case field for which the results are sought. Next, we discuss some of the salient aspects of SIM's *Query Generation* module.
Leveraging Context: SIM creates a search query from the context of the current case, *i.e.,* from the content that is already entered in various fields in the case. In addition, users can provide additional keywords to filter results. A key question that arises is how do we weigh the matches of different case fields in any result in order to arrive at relevance score for the same. We address this design issue with the help of the output of Correspondence Analysis. When we are deriving recommendations for a field β, the query clause created from a field α is boosted by a factor proportional to $Corr(\alpha, \beta)$.
Boosting Relevance: SIM queries are tuned to maximize relevance in several aspects. Firstly, matches in certain fields are weighed more than others (*e.g.,* match in *Title* is more important than a match elsewhere). Secondly, if the different query terms are present close to each other in a document, such a match is weighed more than a match where the terms are far apart. Finally, the relevance score for a result is boosted based on different factors such as users' rating, document age, the number of previous hits on the result, *Quality Score* for the result field and others domain specific rules (*e.g.,* a result is ranked higher if the opportunity was won).

Visualize: One of the key features of SIM is that it delivers readily consumable, summarized information and does not just point to documents from past cases. Figure 4 illustrates how recommendations are visualized in the SIM tool. On the left-hand pane, one finds a list of key topics that are trending in the recommendations generated for a particular case field. A list of topics related to *Win Themes* is shown in the illustration. This list is obtained by clustering the contents of the *Win Themes* field in similar past

[4] http://en.wikipedia.org/wiki/Cosine_similarity

[5] http://en.wikipedia.org/wiki/Euclidean_distance



[redacted] - 05/04/12 - Printer - [redacted]

)Provisioning **cloud** capacity through new C capacity and associated...? To be investiga

Staff Education & Development: i is our responsibility to sustain a capable work force...Implementa replication...Alleviate an estimate capital expenditu

Summaries: Value_Proposition, Risks, Win_Themes, Solution

[redacted] Business Desktop Cloud for [redacted] Customers (SBDC4SC

t Busines S... Solutic

[redacted] Commercial Strategy & Solution Strategy Review [redacted]

Our Value Proposition for Project Printer

Element | Committed Business Outcomes | Solution elements that deliver/enable the committed outcomes

Quality

Fig. 4. Visualizing Results in SIM

cases mined through a contextual search. We use the Lingo algorithm [22] to ensure that the cluster labels are human readable phrases. In the middle pane, we present the recommended case models as well as relevant reference materials. For each result, one can view the *summaries* of different case fields. Also, one may open up a *document preview* for a case field, which shows snapshots of the document regions where the field is documented. Note that such a quick navigation to the desired sections in unstructured documents is possible only because SIM pre-annotates document regions by case semantics. SIM can also provision different forms of visual analytics to further reduce the cognitive load on knowledge workers. For instance, toward the right of Figure 4, we illustrate an example of how one can visually discover interesting associations within case information. The interactive graph visualization shows how different kinds of delivery issues have arisen across different industries.

Discussion: The SIM tool is being actively used in the Opportunity-To-Order process at a large IT Services company and has received positive feedback from its users. They believe that this domain-specific knowledge management system delivers much more precise and contextual results than the enterprise-wide search system they used before. The users suggest that the tool reduces dependencies on personal networks and and yields significant productivity improvements as it jump-starts the case work with relevant information. However, precisely quantifying the improvement in knowledge worker's productivity brought about by SIM's usage on the field is a challenge. Firstly, obtaining usage statistics from baseline systems turns out to be infeasible. Secondly, there are a plethora of factors related to continuous changes in methods, processes and collaboration platforms that need be to be controlled in order to isolate the true impact of the SIM tool. Thus, in the next section, we present a controlled experiment to evaluate the efficacy of a key aspect of our system–generating preferentially weighted contextual queries.

4 Experiments on Contextual Search

In this section, we report experiments conducted to assess the efficacy of our proposed approach in finding the "right" context in order to maximize relevance of recommendations. As discussed in Section 3, when seeking recommendations for a *target* case field, we weigh the query clauses created from the contents of other fields by their respective *Correspondence* scores with the target case field. Our experiments measure the usefulness of such a weighting in improving relevance of recommendations over a baseline approach where equal weights are assigned to each clause in the contextual query.

4.1 Experimental Set-up:

In our experiments, we use a case corpus of 715 case models cataloged from information created during sales engagements at a large IT services company. Each case model was created by aggregating information about a single deal from three different databases within the company. The schema for our integrated case models consisted of 314 fields of different types (*e.g.,* categorical, text, numeric, dates). However, not all case models had all 314 fields; in fact most of them were sparsely populated. For the purposes of our experiments, we choose *Risks* and *Assumptions* as the two target fields for which we generate recommendations. Both of these fields are free text fields; often their contents are organized as a bulleted list of items, sometimes even over a hundred items in a single case. Such lists of *Risks* and *Assumptions* are particularly useful to conduct quality assurance reviews and are a necessary input for crafting clauses in the legal contract when closing a deal.

Competing Approaches: We investigate the efficacy of two approaches of leveraging context in deriving relevant recommendations for a target field. First, we evaluate a *baseline approach* where we construct query clauses out of the contents all non-empty fields in a case model except the target field. In this approach, the matches for all the query clauses are weighed equally while generating recommendations. Second, we apply the *weighted approach*, where we create query clauses from a select set of fields that have a high correspondence score with the target fields. Further, the query clauses are weighed in proportion to the correspondence scores of the respective fields.

Generating Recommendations: We randomly select 8 case models from the case corpus where the fields, *Risks* and *Assumptions* are non-empty. For each case, we construct two queries following the two approaches described above. We execute the queries with Apache Lucene to obtain a ranked list of case models in the corpus with similar contexts. Finally, we retrieve the contents of the target fields in the result case models and present them to an expert who assigns relevance judgments as described below.

Judging Relevance: Judging relevance of a recommendation is hard for anyone who is not actually involved in the case and getting time from case workers to run controlled experiments is always a challenge. However, we manage to work around this issue in the following manner. Note that the case models from which queries were created already have the target fields filled up, so we can use their contents as *ground truth.* Now, the task of comparing two items is much easier than deciding the relevance of a recommendation to a given context. Thus, we ask an expert user who understands the vocabulary of the case domain to compare the recommendation results obtained for a query with its ground truth. The expert chooses one of the 3 labels for each recommendation—*0 : "Not Related"; 1 : "Somewhat Related", 2 : "Related"*. If there is an exact match of any *Risk* or *Assumption* item listed in the recommendation to any item in the ground truth, then it is labeled as "2". If there is some topical match, then the recommendation is marked as "1". Such a labeling strategy helped us collect relevance judgments for each of the top 20 recommendation results derived for the two competing approaches across eight queries for the two chosen target fields; totaling to 640 judgments.

Collecting Metrics: The relevance judgments for the recommendations are used to compute two metrics *Precision@K* and *Normalized Discounted Cumulative Gain (nDCG)*. For computing precision, a relevance label of 0 is considered irrelevant, labels of 1 and 2 are considered relevant. Now, *Precision@K* is defined as the fraction of relevant results for the top-K ranked recommendations. *nDCG* [14] is often used as a measure for evaluating a ranked list with multiple relevance levels. The premise behind Discounted

	Baseline		*Weighted*	
Field	*nDCG*	*P@20*	*nDCG*	*P@20*
Risks	0.504	0.312	0.668	0.456
Assumption	0.57	0.325	0.688	0.49
Overall	0.535	0.318	0.678	0.473

Fig. 5. Summary of Results

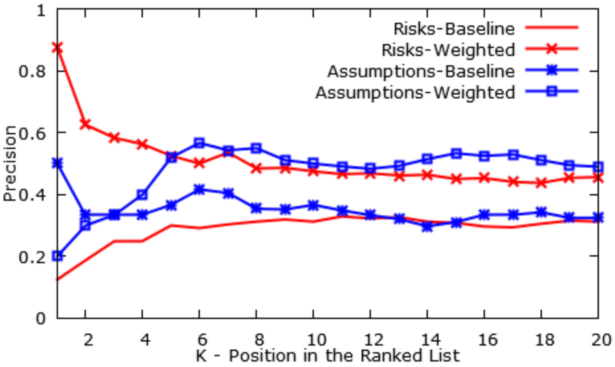


Fig. 6. Precision@K

Cumulative Gain (DCG) is that highly relevant documents appearing lower in a search result list should be penalized as the graded relevance value is reduced logarithmically proportional to the position of the result. For a judgment vector of length p, we compute DCG_p as follows:

$$DCG_p = \sum_{i=1}^{p} \frac{2^{rel_i} - 1}{log_2(i+1)}$$

Next, we re-compute DCG_p after sorting the judgment vector and call it Ideal DCG ($IDCG_p$). Finally, $nDCG$ for the judgment vector is defined as DCG_p expressed as a fraction of $IDCG_p$.

4.2 Experimental Results:

Figure 5 summarizes the metrics *nDCG* and *Precision@20* as observed for the two competing approaches on our dataset. For *Risks*, the *weighted* approach records an improvement of 16.4% in *nDCG* and 14.4% in Precision@20 over the *baseline* approach. For *Assumptions*, we find increases of 11.8% and 16.5% in *nDCG* and Precision@20 respectively. Thus, on average, across the two case fields, both metrics show an improvement of ≈ 15%. Figure 6 plots the values of *Precision@K* for K = 1 through 20 for the two approaches, *baseline* and *weighted*. For the recommendations on *Risks*, the values of *Precision@K* for the *weighted* approach are are consistently higher than those recorded by the *baseline* approach. For *Assumptions*, the curve for the weighted approach is seen to be lagging at K=1,2 but then it surges ahead and thereafter leads the baseline's curve. These results clearly indicate that an intelligent weighting of the different fields in context can improve the relevance of recommendations and that *Correspondence Analysis* can be a viable approach for choosing the weights.

5 Related Work

Recently, there has been a lot of research on Adaptive Case Management and Social BPM technologies for handling ad hoc business processes [30, 20, 29, 28, 17]. However, these efforts have largely focused on developing better case modeling techniques to enhance the level of collaboration between case workers in order to increase the throughput of case processing. Our studies indicate that cases often get stalled because knowledge workers cannot find the necessary information. Thus, we believe that research in case management should attend to the problem of contextually serving information needs of case workers; our work is a first step in this direction.

In the past, reuse of business process information, including formal models and implementation artifacts; has found interest in the BPM community [31]. RepoX [26] and MIT Process Handbook [18] allow storage of business process models with free text search and structured search capabilities. Our past work [12] introduced the notion of contextual search and demonstrated its benefit for requirement gathering activities in SAP engagements. This paper improves upon [12] in the following ways. Unlike the work in [12] that only dealt with textual artifacts, the approach presented in this paper can infer an appropriate weighting of context with different types of fields—text, numeric and categorical. Further, the computation of the strength of an associative relationship between two fields in [12] ignored the influence of other case fields; we address this limitation through the multi-variate modeling in *Correspondence Analysis*. Moreover in [12], the results were not evaluated with the help of relevance judgments from experts. Next, we briefly review related work in other academic communities:

Business Management: Research in Knowledge Management was initiated in the management schools and actively practised in organizations in the professional services industry (*e.g.,* Mc-Kinsey, Ernst & Young). Nonaka [21] proposed a seminal framework to capture the spiraling knowledge processes of interaction between explicit knowledge and tacit knowledge. To bridge the gap between knowledge and information, McDermott [19] proposed to build communities across teams, disciplines, time, space and business units. A comprehensive review of knowledge management systems can be found in [7].

AI & IR: In the mid-1990s, research in Artificial Intelligence (AI) became interested in the problem of solving a new case with the knowledge acquired from solutions to similar cases; the problem was known as *Case Based Reasoning (CBR).* [5] describes a CBR process to represent useful previous experience in a collection of cases, retrieve previous cases similar to a given case, reusing them as a potential solution, revising them by evaluating and adapting the solution, and retaining useful experience as a new learned case for future reuse. Even though CBR techniques have proven to be successful with structured cases, they face issues in scaling to large case corpuses and to unstructured text fields. Limited scale evaluations of hybrid approaches of CBR and textual Information Retrieval (IR) have been reported in the litigation domain [24].

6 Conclusions & Future Work

Most organizations have some mechanisms in place for leveraging prior case knowledge in designing solutions for new cases. It is widely acknowledged that case workers would benefit significantly from reusing the solution patterns that worked in similar past cases. However, finding relevant and contextual information in a timely manner is not always easy within large organizations. Our observations and experience working with a large IT services organization indicates that finding relevant content associated with similar cases is a tedious and manual process. While the approach and the system that we present in this paper addresses some of the relevancy and context related topics, much work still needs to be done. We are currently extending our solution along a number of dimensions to further enrich the knowledge management experience for case workers. Imagine the following query: "show me all applicable offerings, lessons learned, win themes, known risks, known experts, and competitive intelligence related to cloud transformation projects involving migration of SAP CRM application to a private Cloud". Answering such a question requires research in many areas. First, a cognitive system must be able to parse and process a natural language query and identify the component information parts that need to be aggregated. Second, pre-processing of case corpuses needs to be done around well-defined themes associated with cases and information needs to be aggregated around richer collections. This can be done by a combination of clustering and text analysis techniques as the information contains both structured and unstructured content. Graphs need to be built from case corpuses to link related content across cases. Last, an expertise system needs to be created to manage the skills, profiles and experience of case workers so that experts can be recommended for consultation by the case workers. In summary, this is a rich area for research and we are actively pursuing these research topics to further enrich our solution.

References

1. Apache Solr. http://lucene.apache.org/solr/. Accessed: 29/05/2013.
2. Apache Tika. http://tika.apache.org/. Accessed: 29/05/2013.
3. Case Management - Combining Knowledge with Process. http://bit.ly/cErahE. Accessed: 29/05/2013.
4. IBM Content Analytics. http://ibm.co/177lal4. Accessed: 29/05/2013.
5. A. Aamodt and E. Plaza. Case-based reasoning: Foundational issues, methodological variations, and system approaches. *AI Communications*, 7(1):39–59, 1994.
6. C. C. Aggarwal, F. Al-Garawi, and P. S. Yu. Intelligent crawling on the world wide web with arbitrary predicates. In *Proceedings of the 10th international conference on World Wide Web*, pages 96–105. ACM, 2001.
7. M. Alavi and D. E. Leidner. Knowledge management and knowledge management systems: Conceptual foundations and research issues. *MIS Quarterly*, 25(1):107–136, 2001.
8. J. Carbonell and J. Goldstein. The use of mmr, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 335–336. ACM, 1998.
9. S. Chakrabarti, M. Van den Berg, and B. Dom. Focused crawling: A new approach to topic-specific web resource discovery. *Computer Networks*, 31(11):1623–1640, 1999.
10. M. Chui, J. Manyika, J. Bughin, R. Dobbs, C. Roxburgh, H. Sarrazin, G. Sands, and M. Westergren. The social economy: Unlocking value and productivity through social technologies. *McKinsey Global Institute (July 2012)*, 2012.
11. S. Feldman and C. Sherman. *The high cost of not finding information.* Information Today Inc., 2004.
12. M. Gupta, D. Mukherjee, S. Mani, V. S. Sinha, and S. Sinha. Serving information needs in business process consulting. In *Business Process Management*, pages 231–247. Springer, 2011.
13. IDC. *Quantifying Enterprise Search.* 2002.
14. K. Järvelin and J. Kekäläinen. Cumulated gain-based evaluation of ir techniques. *ACM Trans. Inf. Syst.*, 20(4):422–446, Oct. 2002.
15. C. A. Kieliszewski, J. H. Bailey, and J. Blomberg. A service practice approach: People, activities and information in highly collaborative knowledge-based service systems. In P. P. Maglio, J. C. Spohrer, and C. A. Kieliszewski, editors, *Handbook of Service Science.* Springer, 2010.
16. J. Kim, X. Xue, and W. B. Croft. A probabilistic retrieval model for semistructured data. In *Advances in Information Retrieval*, pages 228–239. Springer, 2009.
17. V. Liptchinsky, R. Khazankin, H.-L. Truong, and S. Dustdar. A novel approach to modeling context-aware and social collaboration processes. In *Advanced Information Systems Engineering*, pages 565–580. Springer, 2012.
18. T. W. Malone, K. Crowston, and G. A. Herman. *Organizing business knowledge: the MIT process handbook.* the MIT Press, 2003.
19. R. McDermott. Why information technology inspired but cannot deliver knowledge management. *California Management Review*, 41(4):103–117, 1999.
20. H. R. Motahari-Nezhad, C. Bartolini, S. Graupner, and S. Spence. Adaptive case management in the social enterprise. *Proceedings of the 10th international conference on Service-Oriented Computing*, pg 550–557. Springer-Verlag, 2012.

21. I. Nonaka. A dynamic theory of organizational knowledge creation. *Organization Science*.
22. S. Osiriski, J. Stefanowski, and D. Weiss. Lingo: Search results clustering algorithm based on singular value decomposition. *Proceedings of the IIIS: intelligent information processing and web mining IIPWM*, 4:359–368, 2004.
23. W. Reinhardt, B. Schmidt, P. Sloep, and H. Drachsler. Knowledge Worker Roles and Actions: Results of Two Empirical Studies. *Knowledge and Process Management*, 18(3):150–174, 2011.
24. E. L. Rissland and J. J. Daniels. A hybrid cbr-ir approach to legal information retrieval. In *Proceedings of the Fifth International Conference on Artificial Intelligence and Law (ICAIL)*.
25. S. Robertson, H. Zaragoza, and M. Taylor. Simple bm25 extension to multiple weighted fields. In *Proceedings of the thirteenth ACM international conference on Information and knowledge management*, pages 42–49. ACM, 2004.
26. M. Song, J. A. Miller, and I. B. Arpinar. *Repox: An xml repository for workflow designs and specifications*. PhD thesis, Citeseer, 2001.
27. K. D. Swenson. *Mastering the Unpredictable*. Meghan-Kiffer Press, 2010.
28. K. D. Swenson, N. Palmer, S. Kemsley, et al. *Social BPM*. Future Strategies Inc., 2011.
29. K. D. Swenson, N. Palmer, et al. *How Knowledge Workers Get Things Done*. Future Strategies Inc., 2012.
30. W. M. Van der Aalst, M. Weske, and D. Grünbauer. Case handling: a new paradigm for business process support. *Data & Knowledge Engineering*, 53(2):129–162, 2005.
31. Z. Yan, R. Dijkman, and P. Grefen. Business process model repositories–framework and survey. *Information and Software Technology*, 54(4):380–395, 2012.