# Professional Services Automation: A Knowledge Management approach using LSI and Domain Specific Ontologies

**Vipul Kashyap, Siddhartha Dalal and Cliff Behrens**

Applied Research, Telcordia Technologies
445 South Street
Morristown, NJ 07960
kashyap@research.telcordia.com

**Abstract**

Knowledge Management represents a crucial component of corporations' efforts to gain competitive advantage from targeted re-use of knowledge assets. Companies that provide consulting services depend heavily on developing, selling, and applying ideas. They need to organize and structure intellectual property for efficient and effective re-use. In this paper, we discuss the *RFP Responder*, a prototype system being developed at Telcordia Technologies, for effectively responding to RFPs (Request for Proposals) received by the Professional Services strategic business unit. The use of domain specific ontologies and latent semantic indexing for effective categorization and selection of knowledge assets is presented. The expected business impact of the system is also discussed.

## 1. Introduction

Knowledge is an important resource in many organizations, especially those offering professional consulting services. These companies offer complex non-standardized problem solving services. They differentiate themselves based on their intellectual property (both tacit and explicit knowledge), and technologically advanced team of employees. A high degree of customization is seen where deliverables are tailored towards individual customers [Sveiby 1992; Sveiby 1997].

50% of the fastest growing companies in the US are knowledge intensive organizations that sell the knowledge and know-how of their employees [Mentzas and Apostolou 1998]. Telcordia's Professional Service Strategic Business Unit (PS SBU), provides consulting services for design and deployment of communication networks, and is an example of a knowledge intensive organization. There is a business need for re-use of intellectual property and knowledge assets to compete in the telecom services market. Consequent to this, the PS SBU has launched a KM initiative, to enable effective response to RFPs and opportunities.



We are experimenting with an innovative approach involving the use of latent semantic indexing techniques [Deerwester et. al. 1990] to categorize RFPs and other opportunities wrt to a domain specific ontology of concepts. The concepts capture expertise of business units (BUs) within the PS SBU. The ontology forms the basis of (a) determining relevance of a BU for a given opportunity; and (b) discovery of synergy across BUs.

We first discuss the business problem (Section 2) and describe the *RFP Responder* KM-based framework. An ongoing implementation of the *RFP Responder* is discussed in Section 3. Section 4 discusses the conclusions and likely business impact of the system.

## 2. The Business Problem

The PS SBU at Telcordia Technologies is organized into further sub-units based on the specific technology (e.g., data networking, Cable, DSL) and the market (e.g., CLEC, ILEC, Enterprise) on which they are focused. This sub-division can lead to a "stove pipe" method of operation with little interaction between employees. The following components have been identified to address this problem.

**Responding to RFPs** There is a need for efficient and effective processes that enable PS BUs to respond effectively to large telecom RFPs involving cross-functional work. This involves re-use of work done by other BUs in responding to similar RFPs and opportunities.

**Streamlining Project Planning and Execution** Staffing of projects can be done in a more precise manner by matching skill sets of available employees with the expertise required on a project. Project execution can be accelerated by re-using similar project deliverables developed by other BUs.

We now describe the *RFP Responder*, a KM framework oriented towards re-use of different types of intellectual property. *Relevance determination* is the core functionality to enable effective re-use.

- Determine relevance between RFPs and capability documents describing PS practices. RFPs are obtained

from e-mail, web sites of customers and telecom portals such as Telezoo (http://www.telezoo.com) and Simplexity (http://www.simplexity.com).

- Determine relevance between RFPs and news articles from web sites and trade magazine sites such as TeleCentric (http://telecentric.com), Telezoo and Total Telecom (http://www.totaltele.com).
- Determine relevance between RFPs and previously written proposals to enable proposal re-use.

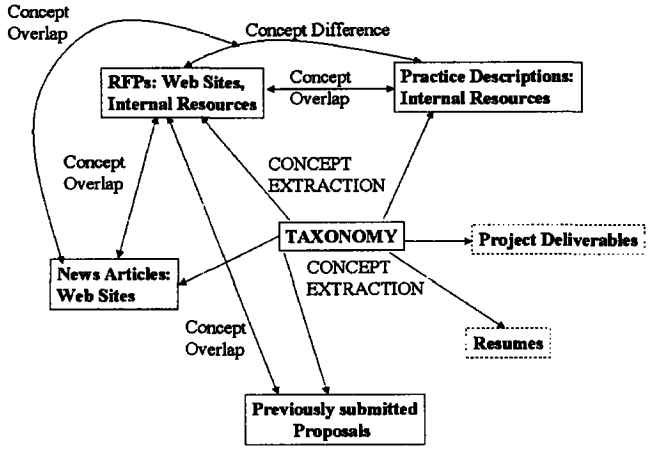



**Figure 1: Components of the RFP Responder**

The components of the RFP Responder framework (Figure 1) are: (a) A *concept taxonomy* describing technology areas and consulting services, which is the central component; (b) techniques for *concept extraction* based on InfoSleuth agent technology [Bayardo et. al. 1997; Hwang 2000] and latent semantic indexing [Deerwester et. al. 1990]; and (c) techniques for *concept overlap computation* for relevance determination. These techniques are being experimented with in the RFP Responder prototype.

## 3. The RFP Responder Prototype

We now discuss in detail various technological components of the *RFP Responder* system (Figure 1). The main components are:

- The *RFP Responder Taxonomy*, the domain specific ontology developed at Telcordia captures technology expertise information in the PS SBU, and forms the core of the system.
- LSI-based techniques for generating a vector space that captures the *latent semantics* embedded in documents containing Telcordia's intellectual property. Concepts in the taxonomy are mapped to coordinates in the vector space.
- Algorithms for concept exaction, which use the mapping of concepts in the vector space to associate a document with various concepts.
- Determination of relevance between opportunities and business units; and discovery of synergy across business units based on concept matching.

### 3.1 The RFP Responder Taxonomy

Keyword-based information retrieval has the major defect of returning large number of hits with a small number of relevant documents. One reason for this is that documents belonging to domains not related to the user query are retrieved. In the *RFP Responder* framework, the search space is restricted to a domain specific taxonomy based on technology areas of expertise. We use LSI as the search engine, which increases precision and recall of the results [Deerwester et. al. 1990; Dumais 1991]. Ontologies and taxonomies have been gaining popularity in areas such as information integration [Mena, Kashyap, Illarramendi, and Sheth 1998; Bayardo et. al. 1997; Arens, Chee, and Knoblock 1993], collaboration between different agents [Gruber 1993], and knowledge management [Benjamins, Fensel, and Perez 1998; O'Leary 1998].

The RFP Responder taxonomy describes technology areas of consulting services. Within a hierarchical taxonomy, the strength of association between a given concept and document also depends on the strength of association between the given document and related concepts in the taxonomy. This association may be via *subclass-of*, *part-of*, and *instance-of* relationships. Each concept has representative vector(s) associated with it. The components and structure of the taxonomy[1] are illustrated in Figure 2.

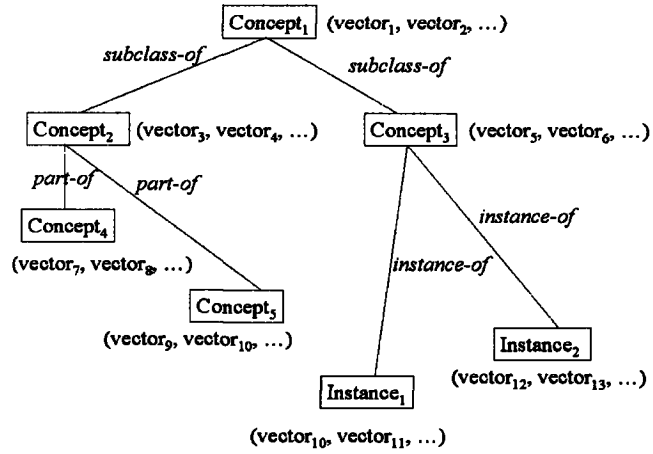



**Figure 2: Structure of the RFP Responder Taxonomy**

### 3.2 Associating an LSI Space with Taxonomic Concepts

We use latent semantic indexing to take advantage of implicit higher order structure in the association of terms

[1] Please note, actual concepts cannot be illustrated due to proprietary restrictions.

with documents ("semantic structure"). We use singular value decomposition, which allows arrangement of term-document associations into a "semantic space" where closely associated terms and documents are placed next to each other. Position in this space serves as a new kind of semantic indexing. We now discuss our approach for aligning a concept taxonomy with an LSI vector space.

3.2.1 Using SVD analysis to generate and LSI vector space
LSI applies singular-value decomposition (SVD) to a term-document matrix where each entry gives the number of times a term appears in a document [Berry, Dumais, and O'Brien 1995]. Typically, a large term-document matrix is decomposed to a set of 200-300 orthogonal factors from which the original matrix can be approximated by linear combination. Roughly speaking, these factors may be thought of as artificial concepts; they represent extracted common meaning components of many different words and documents. Each term or document is then characterized by a vector of weights indicating its strength of association with each of these underlying concepts.

Consider a collection of $m$ documents with $n$ unique terms that, together, form an $n \times m$ sparse matrix $E$ with terms as its rows and the documents as its columns. Each entry in $E$ gives the number of times a term appears in a document. In the usual case, log-entropy weighting *(log(tf+1)entropy)* is applied to these raw frequency counts before applying SVD. The structure attributed to document-document and term-term dependencies is expressed mathematically in the SVD of E:

$E = U(E)\, \Sigma(E)\, V(E)^T$

where $U(E)$ is an $n \times n$ matrix such that $U(E)^T U(E) = I_n$, $\Sigma(E)$ is an $n \times n$ matrix of singular values and $V(E)$ is an $n \times m$ matrix such that $V(E)^T V(E) = I_m$, assuming for simplicity that $E$ has fewer terms than documents. The attraction of SVD is that it can be used to decompose $E$ to a lower dimensional vector space $k$. In this rank-k construction:

$E = U_k(E)\Sigma_k(E)\, V_k(E)^T$

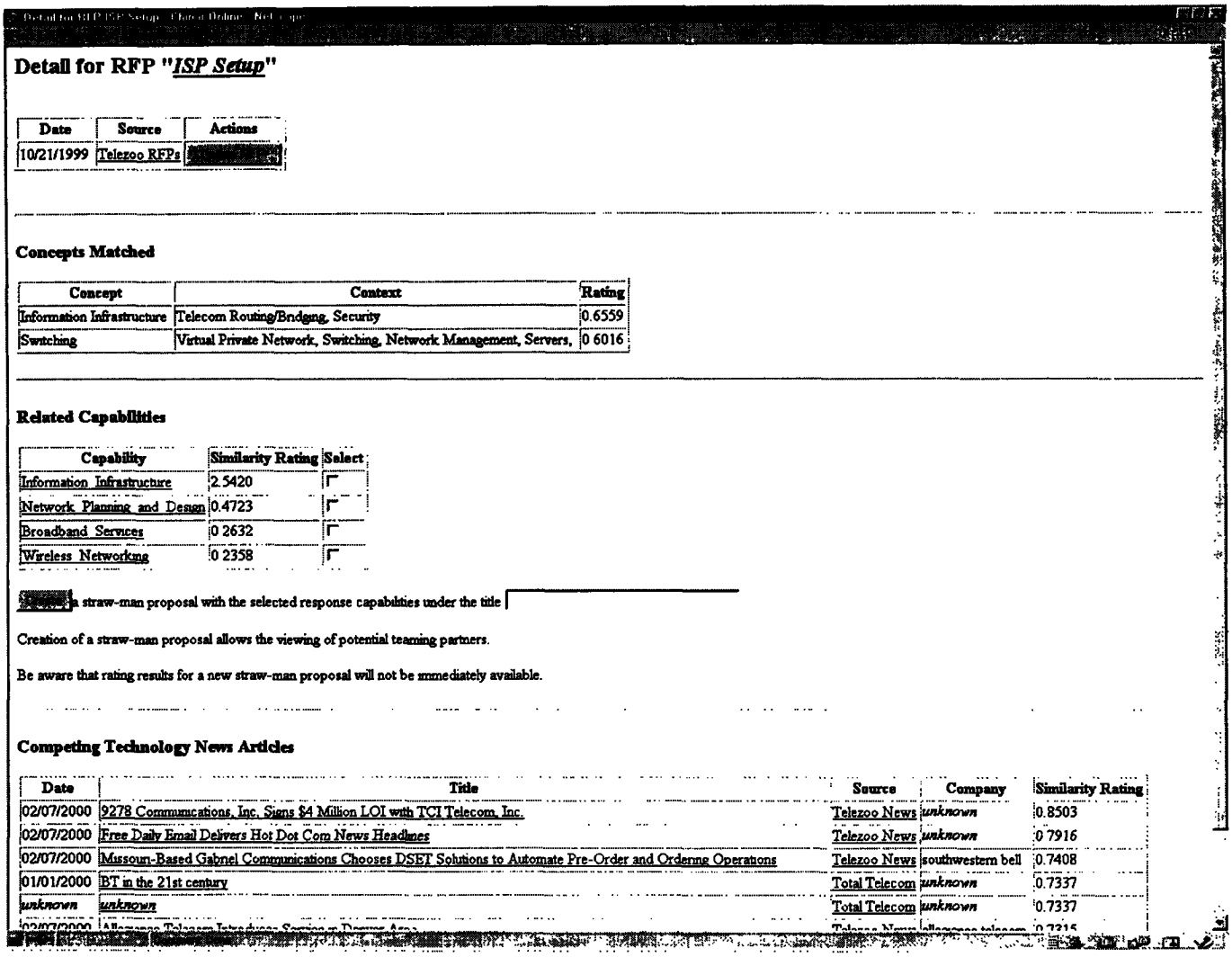
Once the LSI vector space has been constructed, other documents can be "folded-in" to this space, located at the weighted vector sum of their constituent terms. A new term can also be added to this space in an analogous manner by computing its vector from the vectors of all documents that contain the term. In this LSI space, words similar in meaning and documents with similar content, will be located near one another. These dependencies enable one to query documents with terms, but also terms with documents, terms with terms, and documents with other documents. A query is treated as a "pseudo-document", or a weighted vector sum based on the words it contains. The cosine or dot product between term or document vectors corresponds to their estimated similarity, and this measure of similarity can be exploited in interesting ways to query and filter documents, or automatically route them to interested readers [Foltz and Dumais 1992; Dumais 1995]. This measure of correspondence between query vector $q$ and document vector $d$ is given by: $sim(U_k(E)^T q,\ U_k(E)^T d)$. Berry, Dumais and O'Brien [Berry, Dumais, and O'Brien 1995] provide a formal justification for using the matrix of left singular vectors $U_k(E)$ as a vector lexicon.

3.2.2 Creating representative vectors for a concept
For the current application, LSI provides the mechanism for associating a taxonomy with semantic neighborhoods in the LSI-generated vector space. One begins by taking a representative corpus of documents, together with documents that best serve as exemplars, i.e., best capture the meaning, of concepts in the taxonomy. A term-document matrix is constructed for all documents and the LSI space generated. Vectors representing exemplar documents are associated with appropriate concepts in the taxonomy. When a new document is received, the corresponding vector is computed. The cosine similarity measure between this vector and each of the taxonomic vectors determines the best taxonomic classification for the document.

## 3.3 Algorithm for Concept Extraction

We now discuss the algorithm [Hwang 2000] for computing the relevance score of a document *wrt* each concept c ∈ Taxonomy. The set of relevant concepts is identified by filtering through a pre-set threshold. The algorithm consists of the following steps.

- The given document is "folded" into the LSI vector space, and its vector representation in this space is computed.
- The strength of association of a concept $c_i$ with a document is estimated by computing the cosine measure between the vector(s) associated with the concept and the vector representation of the document.
- The measure computed above is incrementally modified based on the association of an ancestor or descendant of $c_i$ with the document. There is greater evidence of a concept document association, if its related concepts (related by the *subclass-of, part-of, instance-of* relationships) are also associated with the document. Also, further the distance of an associated concept, the lesser it contributes to the relevance measure.
- Based on the relevance measure of a document with a concept, the representative vector(s) deemed to be associated with the concept are updated.



**Detail for RFP "*ISP Setup*"**

| Date | Source | Actions |
|---|---|---|
| 10/21/1999 | Telezoo RFPs | [redacted] |

**Concepts Matched**

| Concept | Context | Rating |
|---|---|---|
| Information Infrastructure | Telecom Routing/Bridging, Security | 0.6559 |
| Switching | Virtual Private Network, Switching, Network Management, Servers, | 0 6016 |

**Related Capabilities**

| Capability | Similarity Rating | Select |
|---|---|---|
| Information Infrastructure | 2.5420 | ☐ |
| Network Planning and Design | 0.4723 | ☐ |
| Broadband Services | 0 2632 | ☐ |
| Wireless Networking | 0 2358 | ☐ |

[redacted] a straw-man proposal with the selected response capabilities under the title

Creation of a straw-man proposal allows the viewing of potential teaming partners.

Be aware that rating results for a new straw-man proposal will not be immediately available.

**Competing Technology News Articles**

| Date | Title | Source | Company | Similarity Rating |
|---|---|---|---|---|
| 02/07/2000 | 9278 Communications, Inc. Signs $4 Million LOI with TCI Telecom, Inc. | Telezoo News | *unknown* | 0.8503 |
| 02/07/2000 | Free Daily Email Delivers Hot Dot Com News Headlines | Telezoo News | *unknown* | 0 7916 |
| 02/07/2000 | Missouri-Based Gabriel Communications Chooses DSET Solutions to Automate Pre-Order and Ordering Operations | Telezoo News | southwestern bell | 0.7408 |
| 01/01/2000 | BT in the 21st century | Total Telecom | *unknown* | 0.7337 |
| *unknown* | *unknown* | Total Telecom | *unknown* | 0.7337 |

**Figure 3: A Typical Scenario**

## 3.4 Relevance Determination and Synergy Discovery

The concepts (contained in the taxonomy), extracted from documents form the basis of relevance determination between RFPs, proposals, business units, etc. The results of analyzing an RFP titled *ISP Setup*} are presented in Figure 3. The entries in the table **Concepts Matched** identify the various concepts that have been extracted from the RFP document. The *context* in which they were extracted and the *rating* or the strength of association of a concept with the RFP is also displayed. The entries in the table **Related Capabilities** identify the BUs which have capabilities required in the RFP. The *similarity rating* is a measure of the overlap between concepts appearing in the RFP and BU capability descriptions. Finally, the table **Competing Technology News Articles** displays news articles, which discuss technologies required by the RFP, but expertise about which, is not possessed by Telcordia. The *rating* in this case reflects a measure of *concept difference* as discussed earlier.

**Determination of Relevant RFPs and Proposals** The RFP and BU capability documents, proposals to various RFPs are analyzed to determine the presence of relevant concepts. The concept overlap is computed based on the strength of association of a concept with a document. The concepts *Information Infrastructure* and *Switching* and relevant BUs with related capabilities *Network Planning and Design, Broadband Services* and *Wireless Networks* are identified.

**Discovering Synergy across BUs** The concept overlap between capability documents of various BUs helps identify potential cross-BU collaboration *wrt* an RFP. BUs with capabilities related to *Information Infrastructure* and *Network Planning and Design* have synergies that may be exploited.

**Determining Potential Competitors** News articles from various sources can be analyzed to determine presence of concepts extracted from RFP documents. This enables identification of companies doing related work.

**Determining Potential Collaborators** The difference between the set of concepts appearing in the RFP documents and BU capability descriptions is computed. News articles may be analyzed for concepts in the set difference. This helps identify companies with the required capabilities not possessed by Telcordia.

## 4. Conclusions

Knowledge-intensive organizations need to leverage and re-use their intellectual property in order to maintain their competitive edge in a fast-paced market. The key component is to organize existing information in a way that links structured knowledge (taxonomy) and unstructured information (documents) to enable precise retrieval of relevant information when required.

We have proposed a knowledge management framework, the *RFP Responder* which is being currently developed and deployed within Telcordia Technologies. The key attributes of this framework are: (a) the centrality of a well-designed taxonomy containing domain specific concepts and vectors from an LSI space that enable relevance determination with a high degree of precision; and (b) a high degree of automation in concept extraction leading to smooth integration with the knowledge life cycle.

We expect to achieve significant levels of reduction in the cost and time taken to generate a proposal, and an improvement in the quality and success rate of contract wins. Cost and time reduction in execution of projects by re-use of project deliverables is also expected. A significant increase in cross-organizational synergy is anticipated, which will enable Telcordia to bid on large RFPs. Plans for extending the framework include: tracking of opportunities (RFPs and other "leads"), appropriate design and retrieval of templates for proposal preparation, and a best practices database for speeding up consulting engagements. Relevance determination is expected to be a critical component and will lead to extending the current framework to provide comprehensive support for a consultant on his client engagements.

## References


Arens, Y.; Chee, C.; and Knoblock, C. 1993. Retrieving and Integrating Data from Multiple Information Sources. *International Journal of Cooperative Information Systems (IJCIS), 2(2).*

Bayardo, R.; Bohrer, W.; Brice, R.; Cichocki, A.; Fowler, G.; Helal, A.; Kashyap, V.; Ksiezyk, T.; Martin, G.; Nodine, M.; Rashid, M.; Rusinkiewicz, M.; Shea, R.; Unnikrishnan, C.; Unruh, A.; and Woelk, D.; 1997. Infosleuth: Semantic Integration of Information in Open and Dynamic Environments. In Proceedings of the 1997 ACM International Conference on Management of Data (SIGMOD), Tucson, Arizona.

Benjamins, V. R.; Fensel, D.; and Perez, A. G.; 1998. Knowledge management through ontologies. In Proceedings of the Second International Conference on Practical Aspects of Knowledge Management (PAKM).

Berry, M.; Dumais, S.; and O'Brien, G. 1995. Using Linear Algebra for Intelligent Information Retrieval. *SIAM Review, 35(4).*

Deerwester, S.; Dumais, S.; Furnas, G.; Landauer, T.; and Hashman, R. 1990. Indexing by Latent Semantic Indexing. *Journal of the American Society for Information Science, 41(6).*

Dumais, S. 1995. Using LSI for Information Filtering: TREC-3 experiments. The Third Text Retrieval Conference (TREC-3), National Institute of Standards and Technology Special Publication, pp. 500-525.

Dumais, S. 1991. Improving the retrieval of information from external sources. *Behavior Research Methods, Instruments and Computers, 23(2).*

Foltz, P.; and Dumais, S. 1992. Personalized Information Delivery: An analysis of Information Filtering methods. *Communications of the ACM, 35(12).*

Gruber, T. 1993. Towards principles for the design of ontologies used for knowledge sharing. In Proceedings of the International Workshop on Formal Ontology: Conceptual Analysis and Knowledge Representation.

Hwang, C. 2000. User-centered text summarization based on domain ontologies. Technical Report, SRI-022-00, Micro-electronics and Computer Technology Corporation.

Mena, E.; Kashyap, V.; Illarramendi, A.; and Sheth, A. 1998. Domain Specific Ontologies for Semantic Information Brokering on the Global Information Infrastructure. In Proceedings of the First International Conference on Formal Ontology in Information Systems (FOIS).

Mentzas, G.; and Apostolou, D. 1998. Managing Corporate Knowledge: A comparative analysis of experiences in consulting firms. In Proceedings of the Second International Conference on Practical Aspects of Knowledge Management (PAKM).

O'Leary, D. 1998. Using AI in Knowledge Management: Knowledge Bases and Ontologies. *IEEE Intelligent Systems, 13(3).*

Sveiby, K. E. 1992. The KnowHow Company: Strategy formulation in knowledge intensive industries. *International Review of Strategic Management.*

Sveiby, K. E. 1997. *The New Organizational Wealth: Managing and Measuring Knowledge-Based Assets.* Berrett-Koehler.