# MANAGING CORPORATE MEMORY ON THE SEMANTIC WEB

Nitesh Khilwani[1], J. A. Harding[1*]

[1]Wolfson School of Mechanical and Manufacturing Engineering, Loughborough University

**Abstract**

Corporate memory is the total body of data, information and knowledge required to deliver the strategic aims and objectives of an organization. In the current market, the rapidly increasing volume of unstructured documents in the enterprises has brought the challenge of building an autonomic framework to acquire, represent, learn and maintain corporate memory, and efficiently reason from it to aid in knowledge discovery and reuse. The concept of semantic web is being introduced in the enterprises to structure information in a machine readable way and enhance the understandability of the disparate information. Due to the continual popularity of the semantic web, this paper develops a framework for corporate memory management on the semantic web. The proposed approach gleans information from the documents, converts into a semantic web resource using RDF and RDF Schema and then identifies relations among them using Latent Semantic Analysis (LSA) technique. The efficacy of the proposed approach is demonstrated through empirical experiments conducted on two case studies.

**Keywords**

Corporate Memory Management, Semantic Web, Ontology, Text Mining, RDF (Resource Description Framework), Latent Semantic Analysis

* Corresponding Author: J.A.Harding@lboro.ac.uk

# 1. Introduction

> In an agricultural economy land is the key resource. In an industrial economy natural resources, such as coal and iron ore, and labour are the main resources. A knowledge economy is one in which knowledge is the key resource. (Houghton, Sheehan, 2000)

The current economy, referred to as a service economy, digital economy, or network economy, is emerging as a knowledge and idea-based economy with knowledge and intangible resources as valuable organizational assets (Robinson 2004). The difference with this emerging economy is the magnitude of information and knowledge incorporated into economic activities that has changed the basis of competitive advantage (Fragidis, Paschaloudis et al. 2008). In the current global economy, the key to success is the extent to which business intelligence is utilized, and ideas and innovations are embedded in all sectors of the economy (Robinson 2004). The way a business gathers shares and exploits knowledge is central to its ability to develop successfully. It is concerned with building an insight, foresight and knowledge about their internal management and effectively managing their present and shaping the future.

The knowledge and expertise across the enterprise is arguably the most valuable asset for any organization. This information is like a raw material, which when processed quickly and cost effectively, can be used for value addition. But, if not filtered, refined, delivered and processed properly, the information becomes practically useless (Wick 2001). This observation stresses the requirement for employing knowledge intensive methodologies that facilitate meaningful acquisition, sharing and reuse of information as a main source of competitive advantage for corporations. In organizations, knowledge management is now recognized as an essential strategy for ensuring that the right information is delivered to the right person at the right time.

*Knowledge management* is the process that takes an organization's data and information and turns it into knowledge for tangible benefits for the company. In the context of knowledge management (KM), an explicit and disembodied representation of knowledge is known as *corporate memory* (Van Heijst, Van der Spek et al. 1996). Corporate memory (CM) is the total body of data, information and knowledge required

to deliver the strategic aims and objectives of an organization. It explicitly stores collective knowledge possessed by human capital and technology, organizational structures and organizational culture; and makes it available to the entire company. The major benefits of this explicit storage are: reduced risk of loss of analytical knowledge, better knowledge flow, and reduction in the time and demands placed by human capital (Huang, Tseng et al. 2005, Dadzie, et al, 2009).

Although, knowledge storage and maintenance within a corporate environment have been considered for decades, it has always been an expensive and risky proposition (Lai 2007, Wang et al 2009). It is only with the advent of web technology that information sharing has emerged as a technology area. The Web has enabled firms to make their information available electronically so that users can access and share their resources and expertise electronically. Despite growing interest and efforts, web technology is still primitive in its functionality. Today's web arranges the information syntactically, which means most information has to be interpreted by humans before use, rather than being processed automatically by machines. To overcome this problem, researchers are focusing on semantic web concepts and tools that enable computers to automatically process and understand the information (Casey, Pahl 2003, Lepratti R, 2006). The primary benefit of this new vision is to represent web resources in formalisms that both machines and humans can understand.

Keeping the present technological scenario in mind, this paper aims to:

***Develop a framework for representing corporate memory on the semantic web and identify relations between the semantic documents.***

The idea is to provide explicit semantic descriptions for corporate memory and encode it in an unambiguous and machine readable form. The proposed framework mainly consists of two processing steps: firstly the unstructured text documents present in corporate memory are converted into a semantic web resource using the text mining approach. The TEXT2RDF application was developed to extract information from documents and convert it into a semantic web resource using RDF and RDF Schema.

Subsequently, relations are identified between terms and phrases present within documents and among other semantic documents available in corporate memory, to facilitate effective knowledge reuse. The proposed framework with information extraction and semantic annotation is used to facilitate the following applications in an enterprise (Missikoff, Schiappelli et al. 2003):

- *Semantic Interoperability*: Annotating information resources with local semantics to support business cooperation among enterprise software applications.
- *Document Search and Management*: Assist in navigating and analysing the set of documents by identifying relations between them.
- *Knowledge Management*: Organization and retrieval of enterprise knowledge.
- *Web Services publishing and discovery*: Semantic matchmaking of information producer and information consumer.

The rest of this paper is organized in seven sections. Section 2 details corporate memory and related literature on corporate memory management. Section 3 discusses managing information on the semantic web. In Section 4, a detailed description of the proposed framework is provided. In Section 5, real life applications of the proposed approach are provided. Experimental work is discussed in section 6 and the conclusions are presented in Section 7.

## 2. Corporate Memory: Strategy for Enterprise Knowledge Management

Knowledge and expertise are the most valuable assets for any organization. Effective organizations are those that recognize where these assets reside and identify conditions for fostering their generation and re-use. Quite often, investment in knowledge assets is considered far more profitable than in the physical assets. With knowledge management, organizations seek to create an environment where knowledge is accessible to everyone regardless of their format, requirements or physical locations. Knowledge management is a discipline that provides strategy, process, and technology to share and leverage information and expertise for improved performance, competitive advantage and continuous improvement of the organization. One of the central themes of knowledge

management is the design, building and maintenance of an effective corporate memory (Van Heijst, Van der Spek et al. 1996). A corporate memory is a resource for preserving valuable heritage, learning new things, solving intricate problems, creating core competencies and initiating new situations for both individuals and organizations now and in the future.

### 2.1. *Corporate Memory*

The concept of corporate memory has been defined by several researchers in different ways but one that is widely referenced and appropriate for this research is (Rabarijaona, Dieng et al. 2000):

> Explicit, disembodied, persistent representation of knowledge and information in an organization, in order to facilitate its access and reuse by members of the organization, for their tasks.

Corporate memory is a major asset in any organization. It mainly consists of i) knowledge possessed by people and technology, ii) organizational structure and iii) organizational culture. In order to create and maintain a corporate memory, it is essential to plan and establish a workable record keeping system. According to (Dieng, Corby et al. 1999), the formation of corporate memory relies on the following steps:

- *Detection of needs*: The first and foremost step is to identify the needs of corporate memory, such as who are the users, which tasks they have to perform, in which situations, which knowledge types they need to memorize and retrieve, which tools they use, etc.
- *Construction*: A corporate memory includes all kinds of written documents, such as: reports, notes, newsletters, contracts, licenses etc. The information must follow a useful format to make things easier for others.
- *Diffusion and use*: The corporate memory is useful, if it is well kept and accessible. Therefore, adequate elements of memory must be distributed to the relevant members of the organization.

- *Evaluation*: Corporate memory is expensive and space consuming (computer memory or space) in terms of storage. Therefore its evaluation for appropriate selection is important, from several viewpoints such as economico-financial, socio-organizational and technical.
- *Maintenance and evolution*: In order to get benefits from this knowledge store, it must be maintained and evolved with high priority.

Corporate memory has been a hot topic in management and computer science, at least since the early 1990's. The next section provides a brief review of the published papers on corporate memory.

### 2.2. *Related Work*

Van Heijst et al. presented initial thoughts on how corporate memory should be organized in order to achieve maximum contribution in the competitiveness of an organization (Van Heijst, Van der Spek et al. 1996). Rabarijaona et al. presented advantages of XML meta-language and ontology for corporate knowledge management (Rabarijaona, Dieng et al. 2000). In (Huang, Tseng et al. 2005), XML is implemented to design the structure of knowledge and construct a standard corporate memory according to the characteristics of different data-analysis techniques. An XML transformation process proposed in their paper converted a general document into an XML based document; and this was validated on a manufacturing case study. Verma and Tiwari presented an XML based representation of corporate memory for supplier selection in a global supply chain problem (Verma, Tiwari 2009).

Euzenat developed the Co4 system for building a repository of corporate knowledge with hyper-linked documents providing integration of formal and informal knowledge (Euzenat, 1996). Demian and Fruchter, introduced a prototype system, named CoMem, to support processes of knowledge reuse in a corporate memory, such as finding reusable items and understanding these items in context (Demian and Fruchter, 2006). Vasconcelos et al designed a corporate memory system, named Group Memory System (GMS) using an ontology based model of a domain specific business process and related

individuals and competences (Vasconcelos et al., 2001). Hilbert et al discussed different ways FXPAL are using for automatically creating and retrieving useful corporate memories without any added burden on anyone (Hilbert et al, 2006).

The former research has utilized different web based tools and languages for the representation of corporate memory. The purpose of introducing web technology in corporate memory management is to enhance the effectiveness of communication and management necessary for its capture and reuse. However, even if information is easily accessible it may still not be straight forward to use or fully understand. Current web tools provide no information about the semantics of the described language. It is dependent on the individual programmer or interpreter to understand the semantics of the language from the available description. In order to overcome this drawback, it is necessary to make knowledge explicit and machine interpretable.

In the age of the knowledge economy, corporate memory needs sets of interconnected data and semantic models to communicate and share information. El-Diraby and Zhang 2006 developed a semantic framework for representing and utilizing corporate memory in the building construction domain (El-Diraby, Zhang 2006). This framework is based on a taxonomy with 6000 concepts arranged with explicit definitions and interrelationships. Rios-Alvarado 2009 introduced ODARyT, a semantic web approach to represent and retrieve information in a corporate memory (Rios-alvarado, Medina-ramírez 2009). Chunchen and Jianqiang 2012 proposed a novel rank approach fitting for the enterprise environment using ontologies and prior knowledge about the target domain.

The next section provides details on how information is represented in the semantic web.

## 3. Managing Information on the Semantic Web

It is commonly accepted that the ubiquity of web technology and the internet has initiated and made possible the real time exchange of information on computers. The

concept of the semantic web adds a new level to web technology (Dieng, Corby et al. 1999). The semantic web is envisioned as an extension to the current web technology in which information is annexed with a well defined meaning to enhance the interoperability of computers and people (Casey, Pahl 2003). The declaration of domain knowledge in a machine readable way enhances the understandability of disparate information. The idea is to provide machine processable metadata that describes the semantics of resources and facilitates search, filtering, condensing, or negotiation of knowledge for human users. The basic building blocks needed for representing information on the semantic web are explained as follows:

### 3.1. *Semantic Web Language*

The semantic web relies on ontologies for defining semantics and providing meaning to the data and applications for automatic processing. Ontology is a form of knowledge representation comprising of a set of concepts, axioms, and relationships that describe a domain of interest. With the support of ontology, a user can communicate with the system by developing and sharing a terminology for building knowledge bases for particular domains. Due to its strong implications in conception of reality, it has gained much interest in computational intelligence for defining the basic terms and relationships using the vocabulary of a topic area as well as the rules for combining terms and relations to define extensions to the vocabulary (Lammari, Métais 2004). There are many ontological applications that have been used for information retrieval, knowledge sharing, and natural language processing, and so on. Valaski et al 2012 provides a list of ontological applications in the field of organizational learning, mainly for the sharing of knowledge and formalization.

An ontology is generally developed and used for one or all of the following purposes:

- *Share knowledge*: sharing the structure of information among software agents and people,
- *Reuse knowledge*: reusing it for other systems operating in a similar domain,
- *Make assumption about a domain*: for easier communication and understanding.

In this paper, ontology is used for defining the semantics for web documents and utilizing them for explicit information representation on the semantic web. An ontology can enhance the manageability of large amounts of information resources, including web pages, text documents, email etc. The ontology model developed for representing the web resources is shown in Figure 1. This model mainly represents web documents which have text content to provide information (especially information of an official nature. However, images, audio, video or any other multimedia content is not considered in this model. This model can help academic or industrial collaborators to capture, organize and share information in a unified and explicit manner, without any risk of misinterpretation.

### 3.2. *Semantic Web Grammar*

Informal ontologies hamper the effectiveness of interoperability and lead to ambiguities, delays and unnecessary work. Therefore, ontologies are encoded in a formal language for expressing the concepts in the domain, so that computers can understand and process data easily. Several ontology languages have been developed in recent years and used in the context of the semantic web. XML, a widely known web language in WWW (World Wide Web) community, is the basis for semantic web languages, which include ontology exchange language (XOL), Simple HTML Ontology extension (SHOE) and Ontology markup language (OML) that are the primary extensions of the XML syntax. However, other popular ones are RDF, RDF Schema (RDF-S) and OWL (Web Ontology Language) developed by the WWW consortium (W3C). These languages are an extended version of XML syntax that provides it with greater machine interpretability and semantic interoperability (Khilwani et al, 2009, Mahl and Krikler, 2007). In this paper, RDF is used for defining concepts and making simple relations between them.

RDF is a widely recognized standard for annotating online Web documents and for transforming the HTML Web to the Semantic Web (Davalcu, Vadrevu et al. 2003). RDF was specified by the WWW Consortium (W3C) and was originally designed as a metadata model so that it provides a syntax convention for representing the semantic of data in a standardized interoperable manner. The RDF metadata model is based upon the idea of

making statements about Web resources in the form of "subject-predicate-object" expressions, called triples in RDF terminology. The subject denotes the resource, the predicate denotes the property of the resource and the object expresses a relationship between the subject and the object.

The RDF schema for the proposed ontology (Figure 1) is shown in Figure 2. This schema does not offer anything specific to the proposed framework, except for one key feature - a standard way to describe vocabularies and distribute them over the web.

### 3.3. *Semantic Web Query Language*

SPARQL (SPARQL Protocol and RDF Query Language) is the query language proposed by W3C to query RDF for the semantic web. In syntax, SPARQL resembles SQL to a great extent (with commands like SELECT, WHERE, ORDERBY) but the difference is SQL works on a relational database while SPARQL works on the RDF files. For example, consider a simple document D-1 having a sentence, Sentence 1 with a keyword "knowledge management". A map of this document is shown in Figure 3. A SPARQL query for printing the title of this document with keyword "knowledge management" from possible RDF documents is shown in Figure 4.

Starting from the top, PREFIX is the namespace for the URL defined in RDF. A prefix is simply a short name which can be used instead of typing the full name each time the URL is referenced. For instance, abc:title signifies the URI http://www.yourURL.com/name/1/#title. The SELECT command specifies the variable names which are to be printed in the final result. Note that SPARQL variables are prefixed with either "?" or "$". FROM is used to specify the database. Finally, the WHERE part of the query states the pattern to be extracted from the RDF. In Figure 4, the first pattern in the WHERE part finds the keyword with value "knowledge management" and the second pattern prints the title of the document, respectively.

## 4. Framework for Corporate Memory Management

The corporate memory management is an infrastructure which is guarded and grown to provide technical support for enterprise knowledge management. According to Kuhn and Abecker, corporate memory management is a comprehensive computer system that captures a company's accumulated know-how and other assets to improve the efficiency and effectiveness of activities (Kuhn, Abecker 1997). Linked with knowledge management tasks, corporate memory is used to capture information from various sources of an organisation and make it available for different tasks. D. Grey stated the purpose of corporate memory management as (Grey 2003):

> One of the central themes of KM is the design, building and maintenance of an effective 'corporate memory', a repository, a dare I say it, knowledge-base. Here the intellectual jewels of the organization will reside, easily accessible, expertly indexed, intuitively browsable. Here experts and novices will come for self-help knowledge, they will find the correct solution quickly, be able to apply the solutions with confidence, and learn from the 'collective experience of the organization'.

Considering the above requirements, this paper proposes a framework for managing corporate knowledge in an enterprise. The motivation for the proposed framework is to improve the semantic understanding of corporate knowledge resources. The semantic web is a key technology for achieving better management of knowledge that makes information accessible and understandable not only by humans but also by computers, thus enabling an extended cooperation among people and machines (Berners-Lee, James et al. 2001). In the context of the Semantic Web, there is a meeting point between Web technology and corporate knowledge: both gather heterogeneous and distributed information and share the same concern about the relevance of information retrieval. However, corporate memory has a context, an infrastructure and a scope limited to the organization where they are applied (Rios-Alvarado Ana, Ramírez et al. 2009).

In this paper, a framework is developed to represent and retrieve knowledge from heterogeneous information resources, identify relations among them and utilize it in various enterprise activities such as learning, information interchange, improve information retrieval and information distribution. The proposed framework focuses on

document based information and consists of four steps as shown in Figure 5 and explained as follows:

### 4.1. *Managing Information*

An enterprise of any kind, e.g. an actual company or a collaborative enterprise with distributed companies working together online, deals with the continuous growth of heterogeneous information resources. This information consists of (Huang, Tseng et al. 2005):

- *General information*, described as fundamental information in the corporate memory, for example, subject-oriented information, emails, notes, etc;
- *Technical information*, which includes information (data) analytical technology and information from a data-storage system for example, manuals, guides, application notes, datasheets;
- *Captured information* which is derived from information (data) analytical technology without receiving any content from experts, for example, analytical events, analytical topics, classified information;
- *Refined knowledge* which integrates captured and technical information with subject-oriented expertise, such as feedbacks, reports etc.

The above mentioned corporate information, in general is extracted from data storage systems or domain experts, and may have varying formats and locations e.g. non computational, database-based, document-based, knowledge-based, case-based, Web-based resources. However, the proposed framework focuses on document based corporate memory, distributed through the Web.

### 4.2. *Mining Text Document*

Several kinds of documents can be identified in a document based corporate memory e.g. project documents, technical reports, reference manual, emails etc. The text mining approach, TEXT2RDF, is employed in this paper to extract basic information from the documents available in different formats and convert it into a semantic web resource

using RDF and RDF Schema. The self explanatory Figure 6 shows the overall architecture of TEXT2RDF which is utilized for extracting information from an unstructured text document. This approach includes 5 processes explained as follows:

1. *Document Processing*: The task of pre-processing is to convert these web documents, selected manually or obtained automatically using appropriate search engine technology, into a unified format. . Note that .txt format with single headings and text in separate lines is the required type of input document used in this approach.
2. *Parsing and Filtration*: The task of parsing and filtration is to label terms with corresponding parts of speech (POS) and extract phrases with words/phrases with relevant POS tags for text analysis. It is carried out in 4 steps, as shown in Figure 6-Step 2, i.e.:
   i). *POS Tagging*: Automatic text tagging is the first and foremost step required for higher level analysis of documents in natural language processing applications. In this paper the Stanford NLP Parser is used for POS tagging (Klein, Manning 2003). It is a probabilistic parser with an unlexicalized PCFG (Probabilistic Context Free Grammars) parsing method based on the factored product model. For instance, a tree structure generated from the parser is shown in Figure 7. This parser uses a fixed tag set defined by Penn Treebank[1] to annotate words present in the text, as shown in Figure 7. The POS tagged structure of the sample sentence in shown in figure 7 and different terms and phrases extracted from the text are discussed below.
   ii). *S-P Structuring*: According to linguistic typology, Subject-Predicate (S-P) structure is a basic pattern around which all English sentences are built. In this step, S-P structure is identified from the obtained tree structure.
   iii). *Term and phrase identification*: In this step, key terms and phrases are identified from the S-P structure and stored in a bag of phrases (BOP). BOP is the customized form of bag of words (BOW) with a set of phrases instead of words

[1] http://www.ling.upenn.edu/courses/Fall_2003/ling001/penn_treebank_pos.html

extracted from the document (Nahm, Mooney 2002). The labels used for tagging phrases include, phrase, iPhrase (i.e. important phrases, *defined in this paper*), sentiment, verb, keyword, term and cardinal.

iv). *Stop word filtration*: The BOP extracted in the previous step consists of several POS that are irrelevant for further analysis (e.g. prepositions, determiners, basic english words etc). The task of filtration is to identify relevant information from the identified phrases and pass it for text analysis. Subsequently, the extracted terms and phrases are clustered and labelled with their types.

3. *Text Analysis*: Using the previous steps, the text document is successfully clustered in a bag of phrases. The next task is to tag these phrases with appropriate labels L, where:

$$L = \{Term, Keyword, Phrase, IPhrase, Cardinal, Sentiment, Verb\} \quad 1$$

The rules defined for identifying labels are as follow:

i). *IPhrase*: It is a list of important phrase present in a document, the phrases with keywords and cardinal terms, e.g. found in 1910, based in Armonk, etc. It includes verb phrases with proper noun (NNP) and cardinal numbers in it (Fei et al. 2010).

ii). *Phrase*: It includes all verb phrases with noun phrase (NP), excluding other types of phrases.

iii). *Sentiment*: includes all verb phrases with adjective phrase (ADJP) or adverb phrase (ADVP) in it.

iv). *Verb*: includes verbs (VB) present in the predicate of S-P structure.

v). *Keyword*: contains noun phrases with a proper noun in them.

vi). *Term*: includes all noun phrases excluding keywords.

vii). *Cardinal*: is a list of noun phrases with cardinal numbers (CD) in it.

A simple example explaining the implementation of above rules is shown in Figure 7. These terms and phrases labelled with their types are structured in a formal manner using the ontology model discussed in section 3.

4. *RDF Generation*: As mentioned earlier, an RDF infrastructure is used in this paper for storing documents in the semantic web. In this step, the labelled BOP is encoded into an RDF document using the ontology shown in Figure 1. An example of an RDF file generated for the text mentioned in Figure 7 is provided in **Appendix A**.
5. *RDF Revision*: This is one of the important steps for semantic information management. As discussed previously, a key step in information management is defining the possible users of that knowledge. Therefore, the RDF document is presented to an expert to define the accuracy of the semantic document generated and store it in a well accessible format and place. In order to perform the RDF document revision, different approaches can be followed, e.g. experts going through the documents manually, a web interface can be provided to the experts to facilitate the process of revision, a self-learning based approach can also be used in the web interface, etc. The approach followed for RDF revision depends on the size and number of documents present in the repository. In this paper, a manual approach was followed.

   The above process of text mining is implemented on all the documents in the repository. Subsequently, relations are identified within the documents, as discussed in next section.

### 4.3. *Relating Semantic Documents*

The next task in the proposed framework is to relate semantic documents by identifying relationships between terms and phrases extracted from the documents. In this paper, terms and phrases are handled separately and differently for identifying relations with other phrases in a corporate memory. The approach applied for relating documents in a semantic corporate memory is explained as follows:

### *4.3.1. Keyword and Term*

The hyperlink process in the TEXT2RDF approach is used to link terms and keywords to a single point in a business directory or dictionary. The words labelled as keywords can be selected from the documents available in corporate memory and hyperlinked with a single URI. However, the words tagged as *term* impose severe constraints in finding relations. Organizations often use 1) multiple words that have similar meaning (synonym), and 2) words with more than one meaning (polysemy). These two constraints cause mismatches in the vocabulary used in corporate memory and information retrieval systems. In order to overcome this drawback the following approaches are being developed by researchers to integrate semantic processing into their information storage and retrieval system, e.g. auxiliary structure, local co-occurrence statics and latent semantic analysis (Bradford 2008).

Latent Semantic Analysis (LSA) is one of the most promising tools used for finding patterns of relationships among a collection of documents (Bernard et al 2011). It is an indexing and retrieval method capable of identifying patterns of relationships between concepts, not just matching specific keywords. A key feature of LSA is its ability to extract the conceptual content from text by establishing associations between terms that occur in similar contexts. It has been used in a variety of information retrieval and text processing application such as, text summarization, information discovery, online customer support, information visualization, etc (Bradford 2008). LSA relies on Singular Value Decomposition (SVD) of a matrix (term-document) to determine relationships between terms and concepts used in the documents (Berry et al, 1995). In this paper, LSA is carried out in 4 steps on a set of corporate memory documents and words are labelled as terms in those documents:

1. *Term-Document Matrix*: Let $D = \{d_1, d_2, \dots, d_n\}$ be the set of documents present in corporate memory and $T = \{t_1, t_2, \dots, t_m\}$ be the set of words tagged as terms extracted from the documents. Initially, a term-document matrix $M_{m \times n}$ is

constructed such that $M = [m_{ij}]$ where $i = 1,2,\dots,m$; $j = 1,2,\dots,n$; $m_{ij} = frequency(t_i, d_j)$.

2. *Singular Value Decomposition*: The matrix $M_{m\times n}$ is then subjected to SVD, which decomposes the $M$ into the product of 3 other matrices, such that:

$$M_{m\times n} = USV^* \qquad 2$$

Where, $U_{m\times m}$ = row entities as vectors of derived orthogonal factor values

$V_{n\times n}$ = column entities as vectors of derived orthogonal factor values

$S_{m\times n}$ = diagonal matrix containing scaling values such that when the three matrices are multiplied, $M_{m\times n}$ is reconstructed.

3. *Dimensional Reduction*: In this step, all but the *k* (here *k*=3) largest values are set to 0. This dimensional reduction leads to a matrix $M_k$ such that

$$M_k = U_k S_k V_k^* \qquad 3$$

where, $U_k, S_k, V_k$ have dimensions $m \times k$, $k \times k$ and $k \times n$ dimensions respectively.

4. *Graphing Document-Term*: Finally, the terms and documents are plotted in a $\mathrm{X-Y}$ graph with $\mathrm{U_{m\times k}}$and $\mathrm{V_{k\times n}}$ respectively. Here the matrix values with $\mathrm{k = 2}$ and $\mathrm{k = 3}$ are taken as xy coordinates in the graph. It places both terms and documents in the same graph and thus helps in identifying clusters of terms and documents.

Here is a simple and illustrative example explaining the above LSA approach:

Suppose a set of four text documents, two from financial sector and other two from IT (information technology) sector. In order to demonstrate the analysis step by step, a limited set of keywords is taken from each document. The first step in LSA is to create a word by document matrix. The matrix $M_{7\times 4}$ with 7 keywords and 4 documents is shown in Table 1. In this matrix, each keyword is a row and document is a column. Each cell contains the number of times that word occurred in that title. For example, the word "computer" appears once in F1 and I1 and twice in I2. In general, these matrices tend to be very large and sparse with most cells containing 0.

Once the document-term matrix is built, the SVD is applied for analyzing the matrix. The SVD is used to find a reduced dimensional representation of $M_{7\times4}$ that emphasizes the strongest relationships and throws away the noise. The three decomposed matrices $U$, $S$ and $V^*$is as follows:

$$U = \begin{Bmatrix} -0.631 & 0.336 & -0.689 & -0.121 \\ -0.43 & 0.585 & 0.631 & 0.274 \\ -0.456 & -0.383 & 0.357 & -0.72 \\ -0.458 & -0.631 & 0.002 & 0.626 \end{Bmatrix}$$

$$S = \begin{Bmatrix} 3.881 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 2.43 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1.664 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1.123 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0.987 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0.674 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0.234 \end{Bmatrix}$$

$$V^* = \begin{Bmatrix} -0.325 & 0.276 & -0.827 & -0.215 & -0.265 & 0.124 & -0.037 \\ -0.384 & 0.62 & 0.345 & 0.38 & -0.276 & 0.085 & 0.35 \\ -0.391 & 0.221 & 0.18 & -0.505 & 0.671 & 0.246 & 0.02 \\ -0.509 & -0.039 & 0.181 & 0.053 & -0.118 & -0.416 & -0.719 \\ -0.516 & -0.539 & -0.197 & 0.367 & 0.253 & -0.163 & 0.424 \\ -0.235 & -0.417 & 0.216 & -0.083 & -0.388 & 0.742 & -0.128 \\ -0.117 & -0.158 & 0.215 & -0.641 & -0.418 & -0.409 & 0.403 \end{Bmatrix}$$

Subsequently, the three matrices $U$, $S$ and $V^*$ are dimensionally reduced and first $k$ columns of each matrix are considered for further analysis. In this example, the matrices are reduced to $k = 3$, the first column is thrown and second and third column are used to build the graph. The reason that the first column is not considered is because it correlates with the number of times that word has been used in all documents. It can be considered for analysis by centring the matrix, but it will dramatically increase the memory and computation requirements of the analysis. Therefore, the second and third columns of $U$, $S$ and $V^*$ are considered for building a document-term graph, shown in Figure 8.

The main advantage of this graph is that both terms and documents are plotted in the same graph. Thus, it can be used to identify the cluster of terms and documents closely related to each other. Further, it shows that the terms occurring in any document lie closer in the graph in comparison to the other terms in the set of terms. As evident from Figure 8, the terms lying closer to the document nodes are occurring in matrix $M_{7\times4}$. The distance between documents and terms provided in Table 2 also clarify the closeness of terms with documents. In order to calculate the distance, a Euclidean distance formula is used in this paper.

### *4.3.2. Phrase and IPhrase*

The similarity between two phrases or iphrase is calculated using Dice's coefficient, defined as:

$$Sim(phrase_1, phrase_2) = \frac{2\times CommonTerm(phrase_1, phrase_2)}{NumberOfTerm(phrase_1)+NumberOfTerm(phrase_2)} \qquad 4$$

where, $Sim(phrase_1, phrase_2)$ is the Dice's coefficient (Dice 1945), $NumberofTerm(phrase)$ indicates the total terms in a phrase and $CommonTerm(phrase_1, phrase_2)$ refers to the number of common terms used in $phrase_1$ and $phrase_2$.

### *4.3.3. Verb*

In contrast to the above two approaches used for relating terms and phrases among documents, the verb tags are used for finding relations and clustering sentences in the same document. In linguistics, a verb is considered to be the most important part of sentence linking two nodes (Subject and Object) in a sentence. The terms labelled with verb tags are classified using the predefined classes defined in Verbnet[2]. Verbnet is a hierarchical verb lexicon with syntactic and semantic information for English verbs using Levin verb classes to systemically construct lexical entries (Dang, Kipper et al. 1998). Currently, Verbnet has a list of over 5200 verbs classified in 270 top classes. By using the

[2] http://verbs.colorado.edu/~mpalmer/projects/verbnet.html

Verbnet classes, the sentences present in a document can be easily clustered and classified in different groups. For instances, three groups of verbs used in a sample document can be:

- *Create- create, manufacture, produce, develop, construct, etc*
- *Require- require, need, demand, involve, etc*
- *Use- use, employ, utilize, apply, etc.*

In real time, enterprises can use these verb classes and define their own top level classes for their organizational corporate memory.

#### *4.3.4. Sentiment and Cardinal*

The words labelled as sentiment and cardinal are phrases providing document specific information. Therefore, these terms are only highlighted in the document and stored separately in semantic web resource.

### 4.4. *Querying Corporate Memory*

The semantic corporate knowledge with hyperlinked text and semantic relations with other documents in a corporate memory can be used to expedite various enterprise activities. It can promote knowledge growth, knowledge communication and knowledge preservation in organizations. In order to use this semantic knowledge, knowledge servers can retrieve information from these structured documents using SPARQL query language, discussed earlier in Section 3.3.

## 5. Applications of proposed framework

A basic requirement for an enterprise is that it must be able to share the right information with the right people or department at the right time so that they can work as a single integrated unit whilst keeping their independence and autonomy. As part of an effort to improve cooperation, communication and collaboration, the concept of semantic documents and framework for relating information is proposed in this paper.

This research work has exploited the emerging semantic technologies e.g. ontology modelling and semantic web, text mining to facilitate information sharing and knowledge management in enterprises. In this research, semantic frameworks are developed to convert the enterprise documents into a semantic web resource. These semantic resources can be utilized in different ways, some of which are discussed here.

### 5.1. *Semantic Interoperability:*

The framework proposed in section 4 extracts information from the web documents and populates the BOP. The BOP is further converted into a semantic web resource using RDF and RDF schema. In the BOP, keywords represent unique entities such as the name of a person, organization, place etc; whereas, terms and phrases are used for representing domain specific words and concepts. These terms and keywords often differ from domain to domain and vary according to the enterprise. This hampers the interoperability and sharability of information between machines, people, enterprises etc. In order to overcome this problem, domain specific meaning needs to be added to the terms and keywords available in the document.

The semantic documents created using the proposed framework can be used for tagging the terms and phrases present in documents with their meaning definitions. Generally, enterprises maintain their local classification systems for storing information about terms and terminologies used in their company. These information repositories can be used for adding additional bits of information to the enterprise documents and thus reducing the risk of misinterpretation while document sharing among heterogeneous information systems during the enterprise collaboration. Also a shared repository can be used for adding semantic content in the documents, for example, EUROVOC, a multilingual glossary maintained by the Office for Official Publications of the European Communities (European Union publications office 2005). This glossary contains 6645 concepts from 21 domains with 3636 associative relationships and it is published in 23 languages. Such glossaries can also be used to maintain interoperability and preserve syntactic and semantic content during the exchange of information.

The proposed concept of information extraction and semantic annotation in corporate documents can also support the idea of Linked Data. Linked data, a sub-topic of the Semantic Web, is basically about exposing, sharing and connecting data via URI (Uniform Resource Identifiers) and HTTP (HyperText Transfer Protocol) on the web (Bizer, Heath et al. 2009). The idea is, if enterprises can publish their glossary, directory or other relevant information with Linked Data, then enterprise documents can be linked with this explicit information and converted into a semantic document publishable and accessible over the global information space, without any risk of misinterpretation. This idea can facilitate the sharing of information resources among enterprises from different backgrounds, cultures or locations, or simply enable heterogeneous systems within one enterprise to interoperate at the data level.

### 5.2. *Document Search and Management:*

The semantic annotations and relations can be used to represent text documents in formalisms that both machines and humans can understand and perform intelligent search, querying and reasoning on them. The annotations added in the enterprise documents using local dictionary or published glossaries will create links among the enterprise documents. Similar to Wikipedia pages, the semantic documents can be easily searched and navigated to get the desired information.

The document search can further be improved by glossaries built using an ontology that defines relations among the terms in the glossary. For example, SKOS (Simple Knowledge Organization Schema), a semantic data model built upon semantic RDF and RDFS and used for sharing and linking knowledge organization systems such as thesauri, classification schemes, taxonomies, and any other type of structured controlled vocabularies. The documents tagged using the glossaries built on SKOS will add hidden relations in the documents. For example, consider a glossary having top level term “fruit” and “vegetable” and narrower terms in these categories such as “apple, banana” and “potato, tomato” respectively. Therefore a search result for the term “fruit” will also return documents containing narrower terms.

Although search engines perform this type of relation based search, in a small size document repository the abovementioned semantic search will play a significant role.

### 5.3. *Knowledge Management*

The proposed framework can facilitate the process of knowledge management and improve knowledge sharing within and among enterprises. The legacy systems running at present in the enterprises have not been developed to directly and independently connect with the systems in other enterprises. Moreover, the enterprise information resources are not structured with standard mechanisms, which can be readable and understandable to both humans and machines without any risk of misinterpretation. In order to support the information sharing and knowledge management with less or without human involvement, the proposed frameworks can be used to describe the resources in a way that is understandable and usable by the enterprises.

### 5.4. *Web Services publishing and discovery:*

In an enterprise, web services allow different pieces of software or machines to communicate with each other in a standardized messaging format. It is used as a networking technology to connect computers and devices with each other for exchanging information and combining it in new ways. However, knowledge sharing among multidisciplinary teams in an enterprise is a challenging and complex problem. The two common types of problem that occur in sharing and exchange of information are semantic (same term applied to different concepts) and syntactic problems (different terms used for the same entity). The syntax related problem can be handled by understanding the format of information from producer and requirements from consumer. However, to handle semantic related issues, the proposed framework for semantic documents can be used. The semantic documents will provide common syntax for information sharing and establish common vocabulary.

Enterprises can use building blocks of semantic web (discussed in Section 3) to facilitate the process of information search using web service. In Garcia 2012, SPARQL queries are

used to facilitate the search process on a document repository using semantic web services.

## 6. Experiments

Companies generally maintain a corporate memory to explicitly store the collective knowledge possessed by human capital and technology, organizational structures and organizational culture and make it available to the entire company. A semantic framework is proposed in this paper for corporate memory management on the semantic web. The implementation of the proposed framework on a real life case-study is shown in this section.

### 6.1. *Experimental Setup*

As discussed in Section 2, corporate memory takes into account the diversity of knowledge and information found in an organization. In literature different types of corporate memory have been proposed, e.g. Dieng Kuntz et al. distinguished internal memory from the external memory in an organization (Dieng-Kuntz, Corby et al. 2001). Tourtier proposed four types of corporate memory, i.e. profession, society, individual and project (Tourtier 1995). Based on these findings from the literature, in this paper, two types of corporate memory are considered as a case study, i.e. profession memory and project memory, and the reasons for their selection are explained as follows:

- *Professional Memory*: This captures skills, capability, expertise and competence of that an organization is able to offer. It defines the skills and abilities based on knowledge and experience of its human capital, methods and resources. This information is utilized to coordinate the deployment of a company's assets in ways that give competitive advantage and produce success in the market place. In the past, professional memory was handled by humans having good knowledge about competences and requirements of the company. In the current knowledge economy, it is really difficult to handle such huge amounts of information and thus it is

necessary to model knowledge in a form that is manageable and addressable. In this paper, the proposed approach is used for storing profession memory as a semantic web resource. To demonstrate this implementation, an experimental dataset has been taken from Yahoo Business (http://biz.yahoo.com). The experimental data set consists of brief descriptions of 20 companies from 4 different sectors. The length of summaries varies from 500 to 2000 characters, averaging 1150 characters per description.

- *Project Memory*: This type of corporate memory stores information about projects handled or completed by an organization. Project documents capture the project experience- both good and bad. It is a rich source of knowledge and data for organizations- if organizations have the time and resources to analyze them. In general, these reports are analyzed to get important details and learn lessons from past experience. The results are then used to enhance processes, improve customer relationships, identify specific problem areas etc. However, most organizations lack resources to examine this memory and thus miss important insights thereby leading to a missed opportunity to learn from past projects. The proposed concept of semantic corporate memory is utilized here to extract knowledge from text based reports. To demonstrate this approach, 5 project reports were considered as an experimental data set. These reports vary from 2000 to 10000 characters, averaging 4500 characters per description.

In the next section, these two experimental datasets are analyzed and relations are identified among them (separately) using the corporate memory management framework proposed in paper.

### 6.2. *Semantic Corporate Memory Management*

The corporate memory management framework, explained in Section 4, analyzes a set of enterprise documents by extracting information from their text contents and then identifying relations among the documents. The proposed framework carries out this analysis in four steps, i.e. 1) Converting an enterprise document into plain text format, 2)

Mining information from the text content, 3) Identifying relations among the text content, and 4) Storing these relations for future use.

A Java based web application is coded in this paper to demonstrate the implementation of proposed framework. The TEXT2RDF application developed in this paper is hosted on Google Projects and available here, http://code.google.com/p/text2rdf/. At first, the TEXT2RDF web application is used to extract a list of terms and phrases (BOP) from the text documents. Subsequently, the BOP from a set of documents is pooled together and relations are identified among them. This analysis is carried out separately and differently for the terms and phrases present in BOP. However for brevity and clear illustration of the application, only the terms and keywords present in the BOP are considered in this case study and relations are identified among them.

The latent semantic analysis (LSA) technique, discussed in Section 4.3.1, is used in this paper for identifying relations among the terms and keywords present in BOP. In this analysis, a term-document matrix is created with each cell representing the frequency of the term in a document. Then, the matrix is decomposed using the SVD (singular value decomposition) technique and the result matrices are used for generating the Document-Term graph. The step-wise implementation of the LSA technique on a small-size problem is explained in Section 4.3.1. In this section, only the document-term graphs generated for the two experimental data sets are provided.

A software JAMA[3], Java Matrix Package, is used in this programme for decomposing the document-term matrix using SVD. The graphs are plotted using MATLAB 6.1[4]. The document -term graphs generated from the two experimental datasets are follows:

- *Yahoo Profile*: At first, the terms and keywords were extracted from the documents present in the first dataset having 20 enterprise profiles. The terms and keywords

[3] http://math.nist.gov/javanumerics/jama/

[4] http://www.mathworks.com/products/matlab/

extracted from the documents were pooled together and the abovementioned LSA technique was implemented on them. The document-term graph between 20 documents and 40 terms selected from BOP is shown in Figure 9 and 10. The documents in the graph are labelled with single letters (e.g. I-insurance, T-technology) and words are labelled with their name. After analyzing the graph and measuring the distance of each term from each document, it was observed that terms and keywords are plotted close to their domain specific documents. For example, the term "information technology" is close to the documents from technology domain (labelled as T), whereas the term "stock-market" is closer to the documents from insurance sector. The four clusters are marked in the graph to clearly show that sector specific keywords are placed close to the documents.

- *Project Report*: Similarly, the project reports were analysed by extracting the terms and keywords from the text content and generating a document-term graph from them. However, as the number of terms and keywords extracted from the document were quite high, the distance of each term from the documents were measured and analysed. On analysing the measured distance, it was observed that each document includes mainly three types of terms and keywords, i.e. template specific (T)– present in all documents, company specific (C)- often used in project documents, project specific (P)- used in particular document. For example, the terms like "health and safety" and "contract period" were in all the reports, whereas the keywords like "Phil Taylor" and "South Wales" were in one or few reports. The document-terms graphs generated with these three types of terms and keywords is shown in Figure 11, where documents are labelled as 1-5 and the three types of terms and keywords are labelled as 'C', 'P' and 'T'. These clusters of keywords obtained from the reports can be stored separately and utilized in enterprise knowledge management.

## 7. Conclusion

To achieve and exploit the benefits available from knowledge resources, knowledge management techniques are implemented to share and leverage information and

expertise for improved performance, competitive advantage and continuous improvement of the organization. One of the central themes of knowledge management is the design, building and maintenance of an effective corporate memory. Corporate memory is the total body of data, information and knowledge required to deliver the strategic aims and objectives of an organization. Due to its continual importance and popularity among enterprises, a framework for corporate memory management on the semantic web is proposed in this paper. This framework mainly consists of two processing steps: firstly the unstructured text documents present in corporate memory are converted into a semantic web resource using the proposed TEXT2RDF application. Subsequently, relations are identified among semantic documents available in corporate memory. Further, an exemplar case study is considered to demonstrate the applicability and effectiveness of proposed semantic frameworks.

In this paper, semantic frameworks are developed to convert the abovementioned two types of documents into a semantic web resource. Similarly, other information resources can be considered and advanced concepts can be proposed in the future research. Information retrieval and semantic annotations are mainly the two tools used in the proposed framework. These are considered as the key technologies in the development of Web 3.0, the next generation web. In Web 3.0, the idea is to add meaning to the information on the web and generate an intelligent web by using machine learning, intelligent applications, natural language processing, etc. In this research, the TEXT2RDF application was developed to extract information from the documents and annotate with domain specific meanings for reducing the risk of misinterpretation in enterprise collaboration. In future, this tool can be used in automatic extraction and semantic annotation of information resources, and can also be modified to achieve a higher rate of accuracy and lower computation time.

In this paper, the idea behind developing semantic corporate knowledge was to link well defined meanings with the content of the information resources and utilize them in identifying relations within and among them. However, the semantic framework

developed was limited to only text documents and in future can be advanced for other types of enterprise information resources.

**REFERENCES**


BERNARD, K.; CASSIDY, A.; CLARK, M.; LIU, K.; LOBATON, K.; MCNEILL, D.; BROWN, D., "Identifying and tracking online financial services through web mining and latent semantic indexing," Systems and Information Engineering Design Symposium (SIEDS), 2011 IEEE , vol., no., pp.158,163, 29-29 April 2011

BERNERS-LEE, T., JAMES, H. and ORA, L., 2001. The semantic web. Scientific American Magazine, Available: http://www.scientificamerican.com/article.cfm?id=the-semantic-web

BERRY, M., DUMAIS, S., and LETSCHE, T., "Computation Methods for Intelligent Information Access", Proceedings of the 1995 ACM/IEEE Supercomputing Conference, 1995

BIZER, C., HEATH, T. and BERNERS-LEE, T., 2009. Linked Data - The Story So Far. International Journal on Semantic Web and Information Systems, 5(3), 1-22.

BRADFORD, R., 2008. Why LSI? Latent Semantic Indexing and Information Retrieval, Content Analyst. White Paper edn. Content Analyst Company, LLC.

CASEY, M. and PAHL, C., 2003. Web Components and the Semantic Web. Electronic Notes in Theoretical Computer Science, 82(5), 156-163.

CHUNCHEN Liu; JIANQIANG Li, "Semantic-Based Composite Document Ranking," Semantic Computing (ICSC), 2012 IEEE Sixth International Conference on , vol., no., pp.126,129, 19-21 Sept. 2012

DADZIE, A.-S., BHAGDEV, R., CHAKRAVARTHY, A., CHAPMAN, S., IRIA, J., LANFRANCHI, V., MAGALHÃES, J., PETRELLI, D., CIRAVEGNA, F., 2009, Applying semantic web technologies to knowledge sharing in aerospace engineering, Journal of Intelligent Manufacturing, 20 (5) 611-623.

DAVALCU, H., VADREVU, S., NAGARAJAN, S. and RAMAKRISHNAN, I.V., 2003. OntoMiner: bootstrapping and populating ontologies from domain-specific Web sites. Intelligent Systems, IEEE, 18(5), 24-33.

DEMIAN P. and FRUCHTER R., 2006. A Methodology for Usability Evaluation of Corporate Memory Design Reuse Systems. ASCE Journal of Computing in Civil Engineering, Vol 20(6), pp. 377-389.

DIENG, R., CORBY, O., GIBOIN, A. and RIBIÈRE, M., 1999. Methods and tools for corporate knowledge management. International Journal of Human-Computer Studies, 51(3), 567-598.

DIENG-KUNTZ, R., CORBY, O., GANDON, F., GIBOIN, A., GOLEBIOWSKA, J., MATTA, N. and RIBIÈRE, M., 2001. Methods and tools for knowledge management: A multidisciplinary approach to Knowledge Management. 2nd edn. France: Microsoft Press.

EL-DIRABY, T.E. and ZHANG, J., 2006. A semantic framework to support corporate memory management in building construction. Automation in Construction, 15(4), 504-521.

EUZENAT, J., 1996. Corporate memory through cooperative creation of knowledge bases and hyper-documents, Proc. 10th KAW, Banff (CA), 1996

FEI Xie; XINDONG Wu; XUEGANG Hu, "Keyphrase extraction based on semantic relatedness," Cognitive Informatics (ICCI), 2010 9th IEEE International Conference on , vol., no., pp.308,312, 7-9 July 2010

FRAGIDIS, G., PASCHALOUDIS, D. and TSOURELA, M., 2008. Towards an Educational Model for the Knowledge Economy. Communications of the IBIMA, 3(9), 62-67.

GREY, D., 2003. Corporate memory - the hard way, Knowledge at work, Available: http://denham.typepad.com/km/2003/09/corporate_memor.html.

HILBERT, D., BILLSUS, D., and DENOUE, L., 2006. Seamless Capture and Discovery for Corporate Memory Bookmark and Share The 15th International World Wide Web Conference (WWW2006), May 23

HOUGHTON, J. and SHEEHAN, P., 2000. A Primer on the Knowledge Economy. Centre for Strategic Economic Studies, Victoria, Australia.

HUANG, C., TSENG, T. and KUSIAK, A., 2005. XML-Based Modelling of Corporate Memory. Systems, Man and Cybernetics, Part A: Systems and Humans, IEEE Transactions on, 35(5), 629-640.

KHILWANI, N., HARDING, J.A. and CHOUDHURY, A.K., 2009. Semantic web in manufacturing. Proceedings of the Institution of Mechanical Engineers, Part B: Journal of Engineering Manufacture, 223(7), 905-924.

KUHN, O. and ABECKER, A., 1997. Corporate memories for knowledge management in industrial practice: prospects and challenges. Journal of Universal Computer Science, 3, 929-954.

LAI, L.F., 2007. A knowledge engineering approach to knowledge management. Information Sciences, 177(19), 4072-4094.

LAMMARI, N. and MÉTAIS, E., 2004. Building and maintaining ontologies: a set of algorithms. Data & Knowledge Engineering, 48(2), 155-176.

LEPRATTI, R., 2006, Advanced human–machine system for intelligent manufacturing, Journal of Intelligent Manufacturing, 17 (6) 653-666

MAHL, A., KRIKLER, R., 2007, Approach for a rule based system for capturing and usage of knowledge in the manufacturing industry, Journal of Intelligent Manufacturing, 18 (4) 519-526.

MISSIKOFF, M., SCHIAPPELLI, F. and TAGLINO, F., 2003. A Controlled Language for Semantic Annotation and Interoperability in e-Business Applications, 20-23 October 2003, pp1-6.

NAHM, U. & MOONEY, R., 2002. Text mining with information extraction. In Proceedings of the AAAI 2002 Spring Symposium on Mining Answers from Texts and Knowledge Bases, March 25–2, Stanford University in Palo Alto, California, Publisher: AAAI.

RABARIJAONA, A., DIENG, R., CORBY, O. and OUADDARI, R., 2000. Building and searching an XML-based corporate memory. Intelligent Systems and their Applications, IEEE, 15(3), 56-63.

RIOS-ALVARADO ANA, B., RAMÍREZ, R., CAROLINA, M. and MARCELÍN-JIMÉNEZ, R., 2009. A Semantic Web Approach to Represent and Retrieve Information in a Corporate Memory, OWL: Experiences and Directions, International Workshop, 23-24 October 2009|, Chantilly, Virginia, USA.

ROBINSON, J.P., 2004. What is the new economy. Alabama Cooperative Extension System, 1(4), 1-4.

TOURTIER, P.A., 1995. Analyse préliminaire des métiers et de leurs interactions. GENIE, INRIA-Dassault-Aviation.

VALASKI, J., MALUCELLI, A,. and REINEHR, S., Ontologies application in organizational learning: A literature review, Expert Systems with Applications, Volume 39, Issue 8, 15 June 2012, Pages 7555-7561

VAN HEIJST, G., VAN DER SPEK, R. and KRUIZINGA, E., 1996. Organizing Corporate Memories, 1996, B. Gaines and M. Musen eds pp1-17.

VASCONCELOS J., KIMBLE K., GOUVEIA F., KUDENKO D., 2001. Reasoning in Corporate Memory Systems: A Case Study of Group Competencies. In: Proceedings of the 8th International Symposium on the Management of Industrial and Corporate Knowledge: 243–253

VERMA, A. and TIWARI, M.K., 2009. Role of corporate memory in the global supply chain environment. International Journal of Production Research, 47(19), 5311-5342.

WANG REEN-CHENG, CHANG YAO-CHUNG, CHANG RUAY-SHIUNG, 2009, A semantic service discovery approach for ubiquitous computing, Journal of Intelligent Manufacturing, 20 (3) 327-335.

WICK, C.W., 2001. Teaching an old economy company new economy tricks: knowledge management at a multinational information technology service firm. PhD Thesis edn. Texas: Texas Tech University.

## Appendix A

An RDF file generated for the text mentioned in Figure 6-7.

```
<rdf:RDF
  xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
  xmlns:kmm="http://www.name.com/">
<rdf:Description rdf:about="http://www.name.com/doc#number">
  <kmm:section rdf:parseType="Resource">
    <kmm:sen rdf:parseType="Resource">
      <kmm:id>#0</kmm:id>
      <kmm:svo rdf:parseType="Resource">
        <kmm:s>Internation Business Machines Corporatation (IBM)</kmm:s>
        <kmm:v>develops</kmm:v>
        <kmm:v>manufactures</kmm:v>
        <kmm:o>information technology products and services worldwide.</kmm:o>
      </kmm:svo>
      <kmm:phrase>develops and manufactures information technology products and
services</kmm:phrase>
      <kmm:keyword>IBM</kmm:keyword>
      <kmm:keyword>International Business Machines Corporation</kmm:keyword>
      <kmm:term>information technology products and services</kmm:term>
      <kmm:term>worldwide</kmm:term>
      <kmm:verb>develops</kmm:verb>
      <kmm:verb>manufactures</kmm:verb>
      <kmm:text>International Business Machines Corporation (IBM) develops and
manufactures information technology products and services worldwide.</kmm:text>
    </kmm:sen>

    <kmm:sen rdf:parseType="Resource">
      <kmm:id>#1</kmm:id>
      <kmm:svo rdf:parseType="Resource">
        <kmm:s>IBM</kmm:s>
        <kmm:v>was founded</kmm:v>
        <kmm:o>in 1910.</kmm:o>
      </kmm:svo>
      <kmm:svo rdf:parseType="Resource">
        <kmm:s>IBM</kmm:s>
        <kmm:v>is based in</kmm:v>
        <kmm:o>Armonk, New York</kmm:o>
      </kmm:svo>
      <kmm:iphrase>founded in 1910</kmm:iphrase>
      <kmm:iphrase>based in Armonk, New York</kmm:iphrase>
      <kmm:number>1910</kmm:number>
      <kmm:keyword>IBM</kmm:keyword>
      <kmm:keyword>Armonk</kmm:keyword>
      <kmm:keyword>New York</kmm:keyword>
      <kmm:verb>founded</kmm:verb>
      <kmm:verb>based</kmm:verb>
      <kmm:text>IBM was founded in 1910 and is based in Armonk, New York.</kmm:text>
    </kmm:sen>
  </kmm:section>
</rdf:Description></rdf:RDF>
```

## Figures

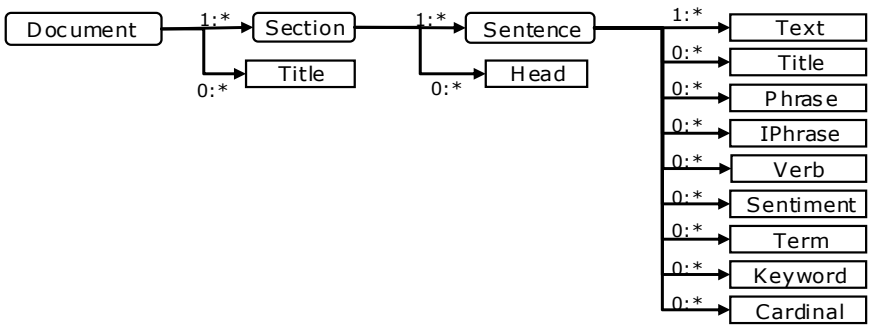


Figure 1: Ontology for Document Model

```
<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
            xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#"
            xmlns:base="http://www.yourURL.com/name/1/">
    <rdfs:Class rdf:ID="Document" />
    <rdfs:Class rdf:ID="Section">
        <rdfs:subClassOf rdf:resource="#Document"/>
    </rdfs:Class>
    <rdfs:Class rdf:ID="Sentence">
        <rdfs:subClassOf rdf:resource="#Section"/>
    </rdfs:Class>
    <rdf:Property rdf:ID="title">
        <rdfs:domain rdf:resource="#Document"/>
        <rdfs:range rdf:resource="http://www.w3.org/1999/02/22-rdf-syntax-ns#Literal"/>
    </rdf:Property>
    <rdf:Property rdf:ID="subject">
        <rdfs:domain rdf:resource="#Sentence"/>
        <rdfs:range rdf:resource="http://www.w3.org/1999/02/22-rdf-syntax-ns#Literal"/>
    </rdf:Property>
    <rdf:Property rdf:ID="subject">
        <rdfs:domain rdf:resource="#Sentence"/>
        <rdfs:range rdf:resource="http://www.w3.org/1999/02/22-rdf-syntax-ns#Literal"/>
    </rdf:Property>
    </rdf:Property><rdf:Property rdf:ID="verb">
        <rdfs:domain rdf:resource="#Sentence"/>
        <rdfs:range rdf:resource="http://www.w3.org/1999/02/22-rdf-syntax-ns#Literal"/>
    </rdf:Property>
    ...
    </rdf:Property><rdf:Property rdf:ID="cardinal">
        <rdfs:domain rdf:resource="#Sentence"/>
        <rdfs:range rdf:resource="http://www.w3.org/1999/02/22-rdf-syntax-ns#Literal"/>
    </rdf:Property>
    ...
</rdf:RDF>
```

Figure 2: RDF Schema

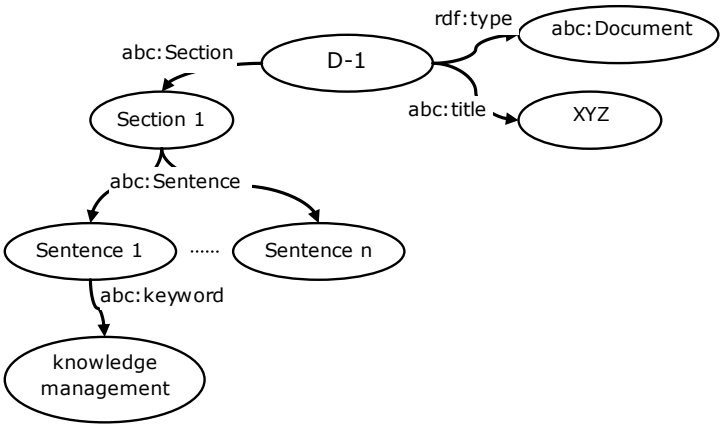


Figure 3: RDF Document Graph

```
PREFIX abc: <http://www.yourURL.com/name/1/#>
SELECT ?Title
FROM <http://www.yourURL/doc/1.rdf>
……
FROM <http://www.yourURL/doc/n.rdf>
WHERE {
  ?x abc:keyword "knowledge management".
  ?x abc:document ?Title;
}
```

Figure 4: SPARQL Query for RDF Graph

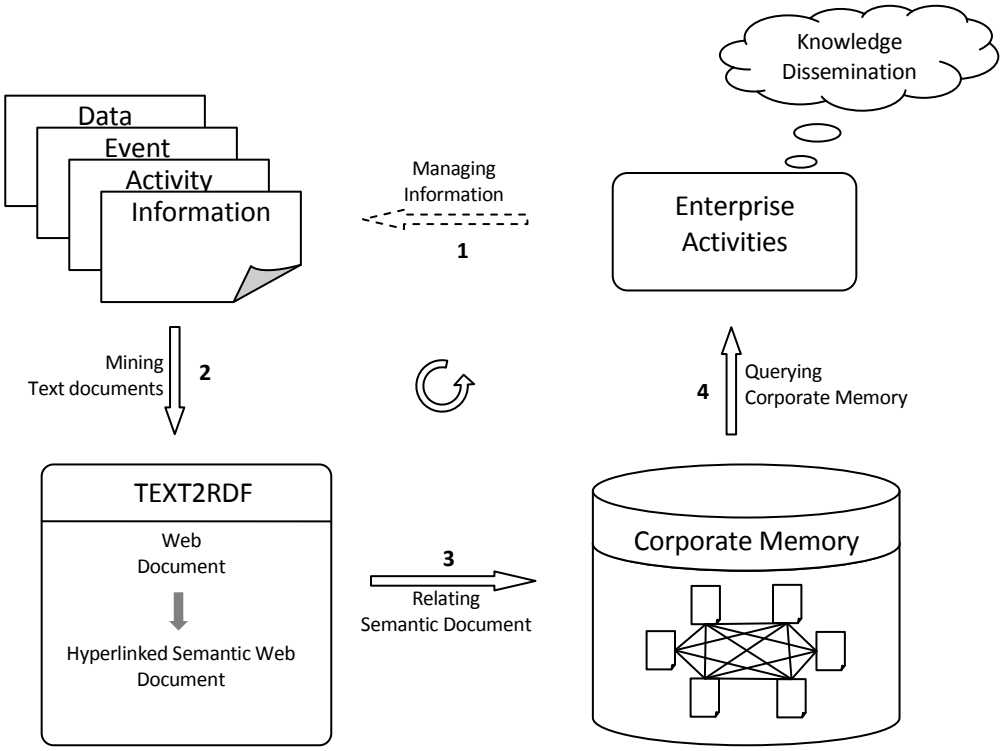


Figure 5: Framework for Corporate Memory Management

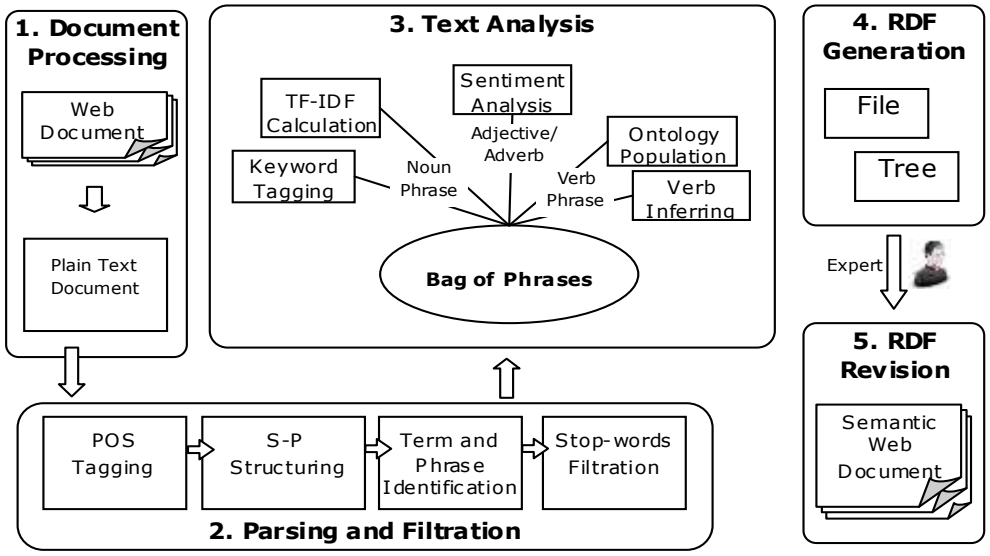


Figure 6: TEXT2RDF Architecture

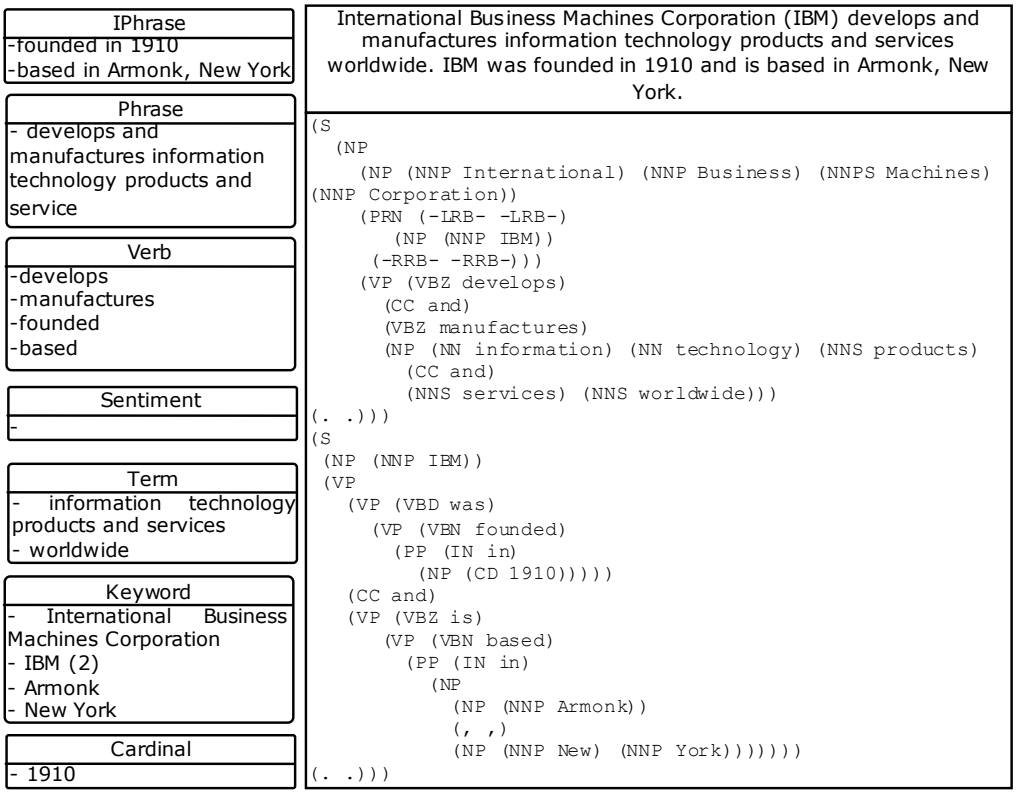


Figure 7: Text Analysis on a Sample Sentence

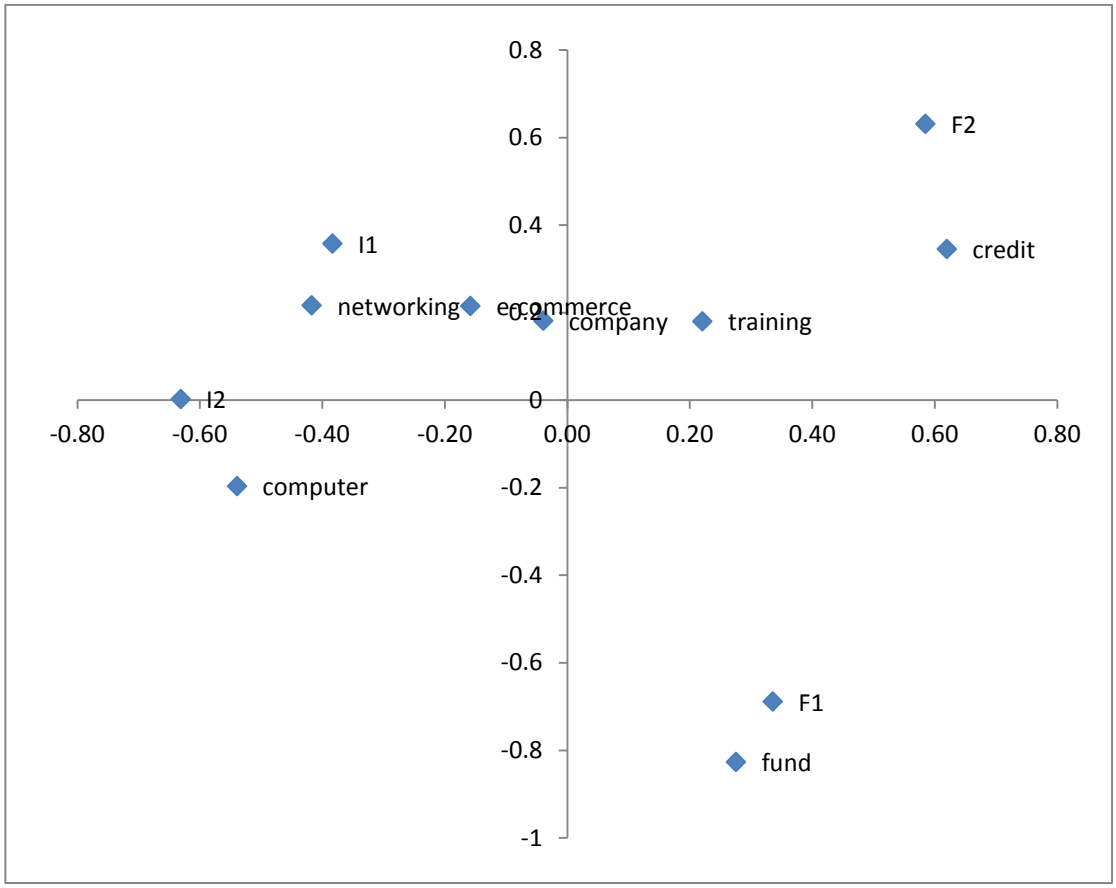


Figure 8: Document Term Graph

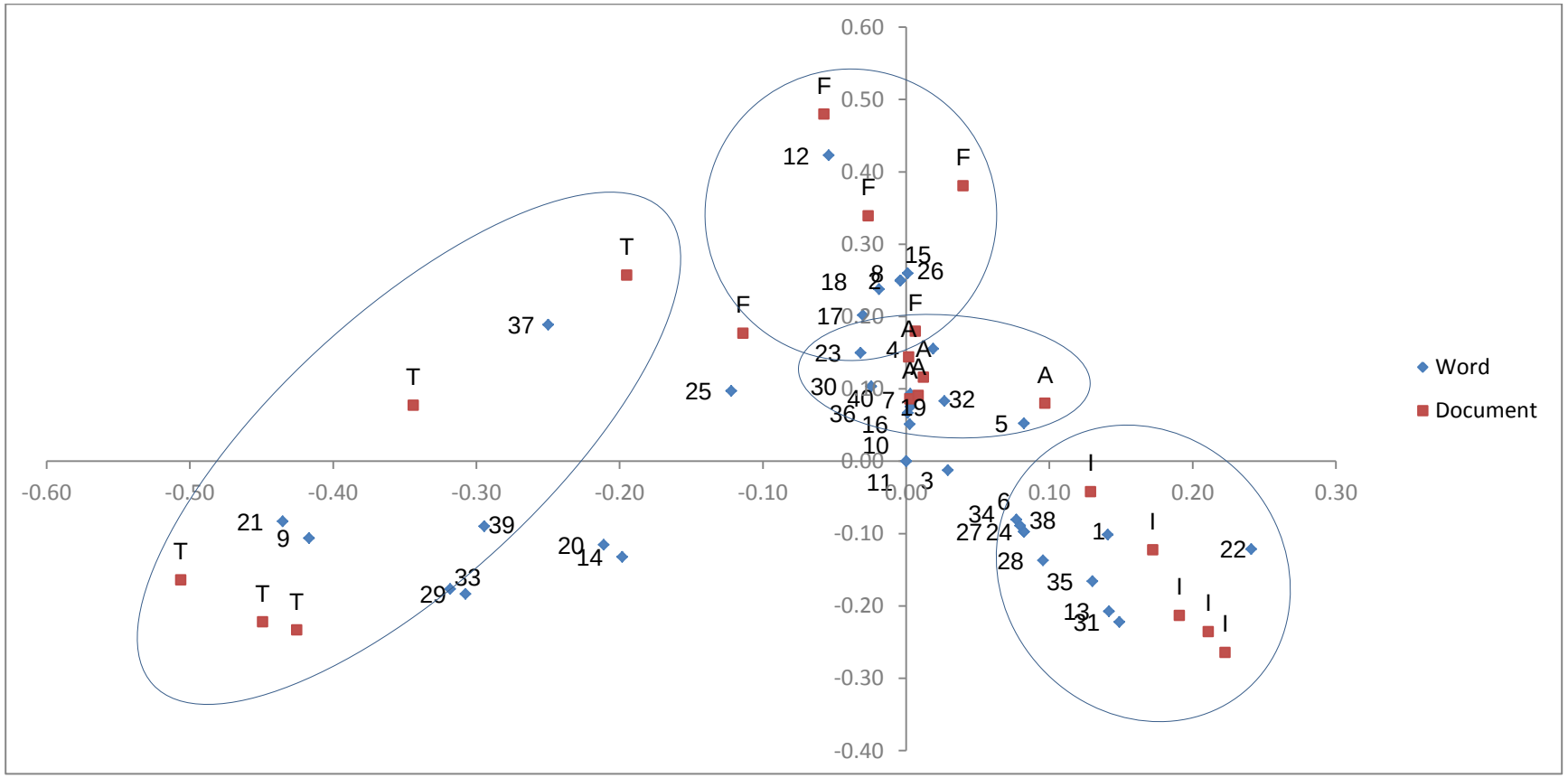


Figure 9: Document-term graph for Enterprise Profiles

Documents: I-Insurance, T-Technology, A-Agriculture, F-Finance.

Words: 1-accident, 2-account, 3-annuity, 4-broker, 5-business, 6-cancer, 7-chemical, 8-company, 9-computer, 10-consulting, 11-credit card, 12-data analysis, 13-disability, 14-e-commerce, 15-fast cash, 16-fertilizer, 17-financial management, 18-fund, 19-greenhouse, 20-hardware, 21-information technology, 22-insurance, 23-interest, 24-investment, 25-knowledge, 26-marketing, 27-medical, 28-mutual funds, 29-network security, 30-poor credit, 31-premium, 32-producer, 33-software, 34-spin off, 35-stock market, 36-sulphuric acid, 37-telecommunication, 38-third party, 39-training, 40-wholesale.


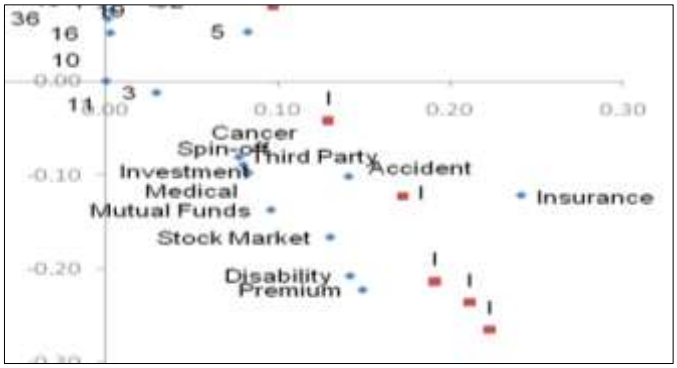


Figure 10: Detailed document-term graph

A: Terms closely related to Documents related to field Technology.

B: Terms related to documents related to field Insurance.

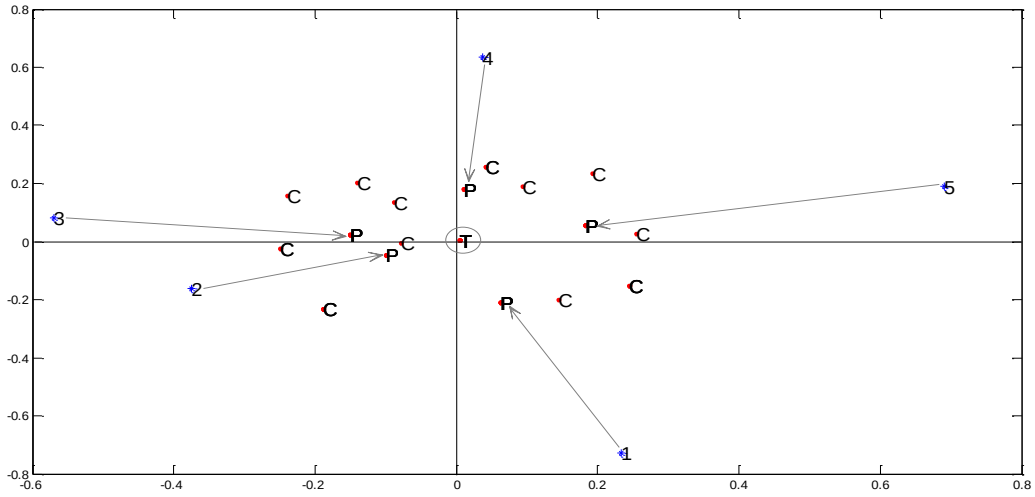


Figure 11: Document Term graph for Project Reports (C- Company Specific, T-Template Specific and P- Project Specific ) (1-5 are Project Report Numbers )

## TABLES

Table 1: Document Term Matrix

| | F1 | F2 | I1 | I2 |
|---|---|---|---|---|
| fund | 2 | 0 | 0 | 0 |
| credit | 1 | 2 | 0 | 0 |
| training | 1 | 1 | 1 | 0 |
| company | 1 | 1 | 1 | 1 |
| computer | 1 | 0 | 1 | 2 |
| networking | 0 | 0 | 1 | 1 |
| e-commerce | 0 | 0 | 1 | 0 |

Table 2: Distance measure between documents and terms

| | F1 | F2 | I1 | I2 |
|---|---|---|---|---|
| fund | 0.022644 | 2.221245 | 1.836137 | 1.836137 |
| credit | 1.149812 | 0.083021 | 1.006153 | 1.006153 |
| training | 0.768386 | 0.335897 | 0.396145 | 0.396145 |
| company | 0.897525 | 0.591876 | 0.149312 | 0.149312 |
| computer | 1.007689 | 1.94896 | 0.331252 | 0.331252 |
| networking | 1.386034 | 1.176229 | 0.021037 | 0.021037 |
| e-commerce | 0.718757 | 0.725105 | 0.070789 | 0.070789 |