SEMANTIC SOCIAL NETWORK ANALYSIS FOR THE ENTERPRISE

Charalampos Chelmis, Hao Wu

Department of Computer Science
University of Southern California
3740 McClintock Avenue, Los Angeles CA 90089-2562
e-mail: {chelmis, hwu732}@usc.edu

Vikram Sorathia, Viktor K. Prasanna

Ming Hsieh Dept. of Electrical Engineering
University of Southern California
3740 McClintock Avenue, Los Angeles CA 90089-2562
e-mail: {vsorathi, prasanna}@usc.edu

Abstract. Business processes are generally fixed and enforced strictly. This is reflected in static nature of underlying software systems and datasets, thereby resulting in resistance to change. However, internal and external situation, organizational changes and various other factors triggers dynamism. Such dynamism is observed in communication channels in the form of issues, complaints, Q&A, opinion, review etc. Channels can be one or mix of email, chat, discussion forum, and official communication, internal social network. Careful and timely analysis and processing of such channels may lead to early detection of emerging trends, critical issues, opportunities, topics of interests, contributors, experts etc. Social network analytics community has been successful in introducing and testing such approaches for general purpose social network platforms like Facebook and Twitter. However, in order to be useful in business context, it is mandatory to integrate underlying business systems, processes and practices with analytics approach. Such integration problem is increasingly recognized as Big Data problem. We argue that Semantic Web technology applied with social network analytics can solve enterprise knowledge management issues while achieving integration.
Keywords: Social networking analysis, Semantic Web, Semantic social network analysis, Information integration, Knowledge networks, Cross-enterprise collaboration, Corporate knowledge management, Collaboration Analytics, Communication Channels, Social search, Knowledge representation, Knowledge acquisition, Collective knowledge, Collective intelligence, User modeling, Ontology engineering,

1 INTRODUCTION

An enterprise’s success is bounded by its capacity for quick adaptation and spontaneity in technological changes and paradigm shifts, rapid response to customer demands through anticipation of future opportunities, and ability to predict, detect and alleviate risk factors and threats. “The competitiveness of firms is related to the adequacy of their decisions, which depends heavily on the quality of available information and their ability to capitalize, enrich and distribute this relevant information to people who will make the right decisions at the right moment” [1].

In a modern enterprise, engineers typically spend 40%-60% of their time seeking information [2, 3]. Same is found true for various other types of employees engaged in different business functions. In such scenario, a system that would enable quick expert identification, facilitate interdisciplinary cooperations that span organizational charts, lessen time spent on searching for solutions, will be pivotal for its success. In high risk operation environments such as smart oilfields [4], shortening the response time required in a failure event, even in the smallest of a fraction of a second, may result in excessive savings, both in terms of environmental and economical consequences.

The end-goal for an enterprise is not storing and managing lots of raw data, but instead, to get to newer actionable business insights faster. We argue that in order for enterprises to get to such insights faster, there is an imperative need for a platform that enables quick, rich, and novel data exploration in multiple, intuitive ways, gleaning information from multiple communication mediums and leveraging it into knowledge. However, due to the way data is generated in a modern enterprise, data management has become increasingly challenging.

1.1 Social Interactions: Big Data - Big Opportunities

Knowledge is generated, captured, utilized and shared without being limited to a specific language or system, but encoded in multiple formats, and distributed over various repositories. In large organizations, knowledge can be handled in the form of standard operating procedures, questioning and answering forums, FAQs, internal websites, social network, personal email correspondence, and other means of communication. In this context, knowledge is observed to be highly dynamic and constantly evolving, and unless otherwise captured, it becomes “buried knowledge” [5].
Due to the richness and variability of systems and tools available in the enterprise information ecosystem, multiple communication channels between employees have become possible. User activity and behavioral data in this context contains valuable information. User’s interests in personal and professional level can be discovered, whereas interesting communication motifs can be mined out, enhancing our understanding on employees’ communication patterns as well as patterns of information propagation and browsing in enterprise networks.

1.1.1 Sources

Enterprises have been mainly relying on e-mail traffic to share information among coworkers [6]. E-mail traffic analysis [7, 8] has lead the extraction of unofficial social networks in enterprise context, in an effort to understand how information flow in the enterprise differs from online social networks. [9] argued that “information extracted from e-mails could prove useful in a knowledge management perspective”, as it would facilitate expert and community identification. Media like SharePoint and Office Communicator are heavily utilized as part of question-answering and problem solving processes, while Active Directory provides a formal structure for employees to comprehend and navigate through the organizational hierarchy, accommodating their need for identifying through browsing, potential collaborators, research teams and business units around the globe, as well as “interesting” projects that others are currently working on.

The wealth of information available in the context of enterprises however, is not limited to formal interactions and silos containing structured data. As social media have become phenomenally popular, enterprises have adopted light-weight tools such as on-line forums and microblogging services for internal communication. Employees have been using social network sites and microblogging services to stay in touch with close colleagues or to reach out to employees they do not know, to connect on a personal level or establish strong professional relationships in order to advance their career within the company [10]. Others, perceive the use of such services as extra communication channel for news reading and company events notification, a mean to promote their ideas, contribute to conversations revolving around company matters, or participate in discussions of work issues.

Enterprise social interactions analysis may lead to various insights, both at a personal (micro) and collective (macro) level. Meaningful micro analysis could revolutionize employees perception of the working environment, offering them better tools for communication, search and productivity, whereas macro analysis could be used for strategic decision making and informed planning.

1.1.2 Macro Analysis

Enterprises can utilize the results stemming out of informal interactions analysis, to better understand how their employees work together to complete tasks or produce innovative ideas, reveal trends, identify experts and influential individuals so
as to evaluate and adjust their management strategy, team building and resource allocation policies.

1.1.3 Micro Analysis

Similarly, employees can benefit in multiple ways. Recommendation services can provide better results in terms of “interesting” people to connect to, as well as suggest “interesting” discussions for employees to contribute to or projects to get involved in. Information filtering algorithms can better promote a subset of news instead of directly delivering all sorts of irrelevant data to employees, alleviating information overload from them, and enabling them to focus on information that does matter. Information acquisition, such as search for people, data and answers to problems can be significantly sped up, resulting in increased productivity through collaboration and problem deduplication.

1.2 Social Big Data Sources: The Big Mess

In this work, we primarily focus on capturing employees’ interests and areas of expertise, as well as mining interconnections between employees’ work-related activities and their social interactions on collaboration platforms used in working environments. In practice, users’ activities are scattered across various collaboration tools used in the enterprise, leaving behind information traces in multiple formats, which is stored in structured, unstructured or semistructured manner. A user might choose chat or microblogging services for casual Q&A sessions, email correspondence for document and ideas sharing, and SharePoint for project tracking purposes. Furthermore, a user may adopt different tools for different projects, or utilize different tools for the same project, depending on current needs. In general, the existence of multiple communication channels scatters information related to a specific employee, establishing the need for an integrated view of users’ activities across platforms.

High-volume activity on enterprise communication channels offers a unique opportunity for content analysis, as active participation of employees and sharing in informal conversations makes it possible to identify knowledge and expertise by analyzing users’ contributed content and its impact to the community. Figure 1 depicts information sources typically found in modern enterprise. Each source offers unique business perspectives, with underlying formats, update frequency, and scope that drives analytics opportunities.

Integrating users’ activities across multiple collaboration tools however is not an easy task. Content from collaboration activities may be significantly short (e.g. 140 characters in Twitter) and inherently noisy. For instance, microblogging content does not adhere to any grammatical or syntactical rules, contains slang terms, user-defined hashtags and emoticons or other special characters, which denote emotions or user-defined notions, the semantics of which may be unknown or not previously modeled. Second, users activities on various collaboration tools signal different kinds of relations, personal or professional [10], of unequal importance [11]. Third, infor-
mation heterogeneity due to different format and schemata or storing mechanism impose further restrictions. Forth, capturing employees’ interests and areas of expertise is a time sensitive task. Existing methods model users’ interests based on static profiles or by keeping track of users’ collaborative activities. However, users’ profiles may be completely unavailable or extremely scarce since users do not often populate enough information to describe themselves, and may be obsolete if users do not constantly update their profiles to match their most up to date interests.

Fig. 1. Big Data in Enterprise

1.3 Summary of Contributions

- We design and build a model that accurately captures the multidimensionality nature of complex, informal, social interactions in the form of orthogonal dimensions. Our model encompasses static enterprise characteristics, as well as dynamic aspects stemming out of collaboration analytics.
- We utilize semantic web techniques for our conceptual modeling and representation. This approach facilitates use of shared domain Ontologies, and Linked Open Data.
- We show that our model facilitates integration, capturing, search, and retrieval of dynamic and constantly evolving enterprise knowledge captured from informal, social interactions.
- We show how our model enhances collaborative data analysis in the enterprise, revealing latent topics, expertise, and interests, both at micro and macro level.
- Our approach leverages knowledge stemming out of informal communication from multiple sources, driving multiple applications, such as team building, re-
recipient recommendation, and event recommendation. We present a case study on a large scale dataset from a Fortune 500 company.

2 REPRESENTATION & ANALYSIS OF BIG SOCIAL DATA

Researchers have modeled, captured and analyzed interactions between people in a plethora of situations [12]. To better mine and understand such complex interactions, their properties and characteristics, the need for some appropriate representation of them has consequently emerged.

Traditionally, researchers have applied graph theory [13] on various graph representations so as to unravel certain features of the network, identify the most important actors in a social network and discover community structures. Social Networks have been historically represented as sociograms [14], in which nodes represent users and arcs represent explicit relationships between them. In order to exploit implicit relationships between users, tripartite graph models, have also been proposed [15, 16].

Social networks evolve when users “friend” each other. However, friend-of links fail to capture the strength of association between users and explicit relationships between them. For example, two users may be computer programmers, but interested in PHP and Java respectively. In this scenario, linking users based on a specific programming language misses the latent relationship in the dimension of computer programming. Typically, social networks capture relationships in a one-dimensional manner: two users are connected by a single edge carrying the generic “friend-of” label.

Rich human interactions and socially generated data can’t be represented using only models of graph theory. In fact, modeling and analyzing social networks with graph theoretic approaches lead to considerable loss of information, ignoring edge semantics. Edges may be temporal and associated to a particular event (e.g. place and time) or may hold for a particular context. Working relationships are often completely disjoint to family or friendship relationships for instance. Further, edge semantics may vary depending on the types of nodes that are connected and the type of interaction between them. The meaning of an edge linking an individual to a document could be modeled for example as “author-of” or “reviewer-of”, depending on the modeler’s intention. Such considerations are partially addressed by edge labeling, which however lacks semantic links to structure them.

Semantic web frameworks answer the problem of representing social data with a rich typed graph model, a query language and schema definition frameworks. Ontologies are used to describe users and their activity, content and its relation to users. Different types of relationships, trust levels and edge weights impose finer grained description using vocabularies. [1] proposed an architecture based on the Semantic Web stack to analyze online social networks while being semantics aware. Its purpose is to explore RDF-based annotated profiles and users’ interactions in

\footnote{http://www.w3.org/RDF/}
Semantic Social Network Analysis for the Enterprise

Social networks using background knowledge (domain vocabulary), predefined ontologies and OntoSNA, an ontology of Social Network Analysis, which provides a way to compute sociometric features using SPARQL\(^2\). This work extends classical graph theory algorithms with semantic features, such as types of resources (e.g. foaf:Person) or properties (e.g. foaf:knows or relationship:worksWith).

Semantic annotation transforms unstructured data into a structured representation that enables applications to better search, analyze, and aggregate information. Twitter users adopted hashtags to alleviate the significant information overload that the streaming nature of social media imposes to users interested in specific topic(s). Hashtags have been exploited for content management, organization and filtering [17, 18, 19]. Even though user-defined hashtags are ambiguous and highly heterogeneous, collaborative structures emerge [20, 21]. [18] makes use of annotated microposts together with background knowledge obtained from Linked Open Data to offer advanced search and organizational capabilities. For example, thanks to semantic links between football and sports, all information mapped only to football can be retrieved in queries regarding sports. Multilayered models, which involve the network between people, the network between concepts they use and links to ontologies modeling such concepts have lately been used [22].

As shown in Figure 1, enterprise social network captures just one aspect of enterprise communication, representing social connectivity among employees. As employees add new friends, join groups, participate in discussion and engage in collaborative activities; the activity stream results in frequent change in underlying graph structure. The graph and user contributed content can be semantically enriched and processed, for link prediction, centrality, and other popular graph analysis techniques. Major part of business related transactions are captured and maintained in enterprise databases. Such structured datasets are frequently updated with massive number of transactions and are typically used for on-line analytical processing (OLAP), business intelligence and reporting applications. Business reports on the other hand capture key summary from enterprise datasets, enumerate financial information, trends, opportunities etc. They are typically generated quarterly in the form of unstructured text, requiring some sort of automated preprocessing and analysis. Additional enterprise data sources include sensor streams, geospatial and multi-media content that exhibit varying update frequencies and formats, requiring a plethora of different techniques for automated processing and analysis. Regardless of the source that generates data, all datasets typically have common references of people, processes, activities, places, measures etc. that establish linkages among them. Adding appropriate annotations in pre-processing steps facilitates integration of such heterogeneous data sources.

---

\(^2\) [http://www.w3.org/TR/rdf-sparql-query/](http://www.w3.org/TR/rdf-sparql-query/)
3 THE BLISS OF MULTIDIMENSIONALITY

Time affects interactions, constituting temporal context (TMC). As time progresses, background knowledge is acquired and learned skill-set incrementally updates, interests and expertise change. The current focus of a specific employee may be completely different than what is stated in an outdated personal webpage or CV. Further, social interactions are in many occasions bounded by, at least some, temporal and localization constraints. This refers to spatial context (SC). For instance, face-to-face interactions may only occur when individuals are physically located at the same place at the same time. Extended interactions due to office adjacency or limited communication at a conference further introduce the concept of time sensitive informal communication importance. Participation in meetings, talk, training, conference constitutes event context (EC). Lengthy discussions on a daily basis indicate stronger bond than periodic hourly meetings, which in turn indicate more significance than a sporadic discussion. The relationship between two individuals therefore becomes a function of time and can be explored only as such. Static social networks ignore such interactions establishing an edge between two users if at least some type of communication has happened at least once. We argue that temporal correlations and causal effects between node features and social connectedness can only be manifested and magnified when considered as a function of time.

Employee interests, skills and expertise can change depending on time, work orientation and responsibilities, project focus and overall team competence. From employee’s perspective, interest, expertise, curiosity, familiarity for topics constitutes participation context (PPC). Topics constitute topic context (TPC). Given a context (e.g. a group discussion versus a status update) may yield significant, different aspects of employees’ focus. Depending on personal or professional nature of content, different interests can be mined and different expertise levels can be identified, for disjoint set of topics. Moreover, employees often assume multiple roles in multiple projects (e.g. an employee might act as manager in one project, while being a software developer in another). This can be captured as project context (PRC). Roles and positions, and reporting hierarchy is captured in organizational context (OC). One context can be closely related to one or more other contexts. For instance, employee interests, skills and expertise can differ at multiple points in time, and be different at the same point in time yet within the boundaries of correspondence with different individuals.

We argue that each kind of context can be complex, thus being decomposable to “sub-contexts”. In order to process enterprise communication effectively, it is imperative to establish a comprehensive model of contexts, and semantic links to structure them. Figure 2 shows various contexts in enterprise informal interactions. Collaborative analytics may result in various representations of captured data, providing insights that differ depending on context (e.g. point of view). Such points of views can be considered as dimensions. Dimensions can be defined in the form of a scale (e.g. for time), geolocation, hierarchy (topics, organizational structure) or can be nominals.
### 3.1 Enterprise Contextual Social Interactions

Our comprehensive list of contexts and associated dimensions enables the study and analysis of enterprise informal communication from multiple perspectives. Figure 3 depicts a scenario of enterprise social interactions with context identified. Employee_AABF participates in a project (project context) along with his colleagues. In performing his role, he comes across a problem and posts a question (activity stream context) at on the enterprise social networking platform, where Employee_AAAD and Employee_AADC read the question (activity stream context). Employee_AADC is interested in the topic of the question (topic context) and starts following the question (activity stream context), while Employee_AAAD responds to the question (participation context) using a reference to a Sharepoint page, which was contributed by Employee_AAGG (domain context).

Figure 4 depicts information sources typically used in an enterprise. In order to find right information, employees are forced to deploy elaborate browsing strategies through a combination of multiple information mediums. The wealth of information available across mediums may become overwhelming if not properly processed, stored and presented to the end user in an intuitive manner. However, mining available information sources and considering them in conjunction, might lead to correlations and causality relations that would not otherwise become obvious. We argue that fusing available information sources is imperative for complex analysis of enterprise social communication data.

Context establishes relationships across various activities and artifacts observed in enterprise communication platforms. Once processed and annotated with appropriate contexts, activities can be retrieved as part of advanced search capability.

---

#### Dimensions in Enterprise Social Network Analysis

<table>
<thead>
<tr>
<th>Topic Context (TPC)</th>
<th>Activity Stream Context (ASC)</th>
<th>Temporal Context (TMC)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Topic 1</td>
<td>Activity Stream</td>
<td>Seconds, Minutes, Hours, Days, Weeks, Months</td>
</tr>
<tr>
<td>Subtopic a</td>
<td>Share</td>
<td></td>
</tr>
<tr>
<td>Subsubtopic a.i</td>
<td>Link</td>
<td></td>
</tr>
<tr>
<td>Subtopic b</td>
<td>Question</td>
<td></td>
</tr>
<tr>
<td>Subsubtopic b.i</td>
<td>Review</td>
<td></td>
</tr>
<tr>
<td>Topic 2</td>
<td>Recommend</td>
<td></td>
</tr>
<tr>
<td>Topic 3</td>
<td>Like</td>
<td></td>
</tr>
<tr>
<td>Topic 4</td>
<td>Opinion</td>
<td></td>
</tr>
<tr>
<td>Project Context (PRC)</td>
<td>Participation</td>
<td></td>
</tr>
<tr>
<td>Event Context (EC)</td>
<td>Interest</td>
<td></td>
</tr>
<tr>
<td>Participation Context (PPC)</td>
<td>Expertise</td>
<td></td>
</tr>
<tr>
<td>Domain Context (DC)</td>
<td>Curiosity</td>
<td></td>
</tr>
<tr>
<td>Organization Context (OC)</td>
<td>Familiarity</td>
<td></td>
</tr>
</tbody>
</table>

---

**Fig. 2.** Contexts in Enterprise Social Network
Figure 5 represents two of such advance search scenarios that lead to macro and micro analysis of social network content. Lens capability permits identification of a users’ activity in specific communication channel (e.g. Employee AAAD’s contribution in Q&A forum), thereby revealing the nature and impact of contribution. We call this capability “lens” due to its focus on a specific, narrow aspect defined by a context. However, a context can be broken down into sub-contexts in a hierar-
chical fashion. Topic and location for example can be expanded to cover subtopics and smaller location segments respectively. Query results in such case constitute a “canopy”.

![Fig. 5. Lens and Canopy Search](image)

### 4 RESONATE: SEMANTIC SOCIAL NETWORK ANALYSIS FOR THE ENTERPRISE

We propose a formal modeling that abstracts the semantics of informal communication into an integrated, context aware, time sensitive, multi-dimensional space, enabling the correlation of seemingly different domains so as to investigate them in conjunction. We introduce a novel social graph representation, shown in Figure 6, which not only contains social links between users but also maintains integrated information regarding users dynamically changing interests and activities, throughout collaboration tools used in working environments.

**Social Layer** captures users contextual and temporal interactions. Nodes represent users and arcs represent explicit relationships (links) between them. To construct the social layer we start with friendship relationships from the social network and augment this initial graph by mining user relationships out of available information sources: e-mail correspondence, chat networks, blogging activity, shared bookmarks and common ratings. An edge between users is defined by the context under which it is created and has an associated timestamp. A user $u$ may be connected to user $v$ under multiple contexts (e.g. sending e-mails and social status updates) in multiple time instances.

**Content Layer** captures published content from all available sources, including but not limited to resources shared by users (e.g. photos or videos), bookmarked
and/or tagged resources (e.g. URLs), users’ generated content (e.g. status updates in Facebook), e-mails, chat messages, and blog posts. Depending on available computational resources or application need, this layer may maintain raw content, which is meant to be processed later on, or in the other extreme only contain the aggregated post-processing results of previous analysis. In the latter case, provenance metadata are to be maintained in unison with analysis results so as to describe for instance the procedure followed and data sources used in the analysis.

Semantic Layer contains meta-information about content, and can be broken into several constituting layers, each containing different metadata about content. This layer may include, but is not limited to, domain ontologies, vocabularies, and folksonomies and taxonomies, external sources of formal knowledge, and linked open data. OpenCalais\(^3\), AlchemyAPI\(^4\), and Evri\(^5\), WordNet\(^6\), and Freebase\(^7\) are examples of semantic information providers and annotation enablers, exposing rich APIs for text analysis and text annotation, entity identification, and topic discovery, as well as complex relationships mining. Linked Open Data\(^8\) can further be exploited to gain insights into knowledge that may not be inherently present in the system under examination, but is accessible through external sources.

![Fig. 6. Layers](image)

We call our framework **Semantic Social Network Analysis for the Enterprise** (rESONAtE). rESONAtE enables analysis which spans layers, considering both

---

\(^3\) [http://www.opencalais.com/](http://www.opencalais.com/)

\(^4\) [http://www.alchemyapi.com/](http://www.alchemyapi.com/)

\(^5\) [http://www.evri.com/](http://www.evri.com/)

\(^6\) [http://wordnet.princeton.edu/](http://wordnet.princeton.edu/)

\(^7\) [http://www.freebase.com/](http://www.freebase.com/)

\(^8\) [http://linkeddata.org/](http://linkeddata.org/)
multifaceted data and metadata, and the underlying informal communication graph. Knowledge is discovered, captured and inferred based on such complex information. In previous work [23] we define rigorous social metrics, which we use to calculate semantic similarity scores between any two object types, users and content alike, in a joint semantic space, given a context. Next, we explain how we materialize our framework and associated metrics to intricately model and extensively analyze a real world dataset from a Fortune 500 multinational company.

We instantiate rESONAtE in the form of Ontology. The main reason we use Ontology, is that it provides a generic, reusable, and machine understandable model for representing the concepts and properties required for describing user activities and measuring their behavior. Figure 7 depicts the coverage of rESONAtE Ontology.

Figure 8 shows an overview of classes in our ontology, which defines a structured set of concepts to characterize employees in comprehensive dimensions.

4.1 Static Modeling

Business_Associate class constitutes a high level abstraction of Employee and Organization. An employee may have various types of connections, specified for multiple contexts. For instance, an employee may have multiple supervisors while being assigned to multiple projects, and at the same time maintain a list of email contacts. An organization provides employment to workers and interacts with (e.g. sells products) other organizations. Treating Employee and Organization as direct descendants of Business_Associate class provides a mechanism to leverage atomic features to a collaborative level. Further, employees can interact (both explicitly and
implicitly) with organizations, while organizations may participate in discussions.

Organization constitutes of Companies that contain various Departments and Facilities. Each Employee holds a Position in the enterprise and participates in one or multiple Projects. Unlike user profiling [24, 25], which mainly focus on static properties such as personal information, we emphasize on the usage of our ontology in dynamic enterprise context. Projects that each employee participates in record the working status of the employee. The “StartDate” and “EndDate” describe the life span of the Project, based on which we can trace projects status. Moreover, it is possible for an employee to participate in multiple projects simultaneously.

Position class describes the current status of an Employee in the enterprise and reflects on current responsibilities and focus area. It contains information such as the Department and Company for which an employee works, and in which Facility she is currently located. The exact employee position title (e.g., manager, senior engineer or CEO) is linked by “hasType” to the object class Employee Position. To represent the hierarchical level of the employee in the company, we use “Orga-
nizationLevel”. “EffectiveDate” denotes the starting date of an employee work at a position. “ExpiryDate” specifies the date that a position ends. Based on the timeline of Positions an employee has worked at, the complete working history of the employee can be traced back, enabling features like future position recommendation.

Accumulated skills and past experiences can be summarized from employee resumes and profiles from past appointments, even though it is difficult to measure employees expertise objectively for different domains [26]. Previous positions and projects provide hints on employees’ specialization areas, however, peer validation in the form of informal interaction acts as supporting evidence of the level of expertise an individual possesses. Expertise class is used to represent levels of expertise for each person with respect to various areas. We use the concept Topic (e.g., computer technology, management, accounting) to represent areas of expertise, as well as Interests. Each Topic has a list of Keywords with probabilities that represent the extent to which a keyword conveys the semantic meaning of a topic.

Connections capture relationships between Business_Associate instances (i.e. Employee or Organization objects). Each connection is established under a specific context. Context can be defined as any information that can be used to characterize the situation of an entity [27]. Context can be a physical property such as time and location, or a logical concept such as a situation [28]. In our model, we consider two kinds of context: Connection_Context and Content_Context. Connection_Context specifies the kind of relationship a pair of entities has, such as colleagues, collaborators, or dominance/subordinate. Content_Context represents the common Interest two entities share. For example, in enterprise social media, an employee can be connected to others with follower/followee relationships. Two employees may share common interest on a Topic such is computer technology. Since context is a function of time, our Context class contains a timestamp property, which is used to record the period of time over which a context is valid.

4.2 Dynamic Modeling

We mine content of informal interactions between employees to capture their expertise in a latent topic space. Recent studies in machine learning area have developed probabilistic models to automatically uncover latent “topics” in natural language texts. Topic Models [29, 30] take advantage of co-occurrence of words in text documents. They use hierarchical Bayesian models to capture the generation process of words in documents by introducing an intermediate latent topic layer. Topic models can address problems of synonymy and polysemy in natural language processing. Each topic is represented as a mixture of words with probability distribution and document can be decomposed into a distribution of various latent topics.

In our work, we adopt the Author-Recipient Topic model (ART) [31] to discover topics in messages posted by employees in internal social media. The model builds on Latent Dirichlet Allocation [29] and the Author-Topic model [32]. Instead of only modeling topic distributions over messages or authors, the ART model conditions the distribution over topics on both the sender and the recipient of a message.
Therefore, latent topics are discovered according to relationships between people. Furthermore, using ART we are able to identify not only how often employee pairs interact, but also which topics are the more prevalent in their discussions.

ART makes the assumption that each word \( w \) in a message is sampled from a multinomial distribution \( \phi_t \) (the word mixture for a topic \( t \)). The topic is drawn from a multinomial distribution \( \theta_{ij} \), which is the topic mixture specific to the author-recipient pair \((i, j)\) of the message. We train ART using the default hyperparameters \((\alpha = 50/T, \beta = 200/V)\), and \( T = 100 \) topics. Inference of ART is achieved by using 1000-iteration Gibbs sampling [31]. We use the trained model to capture not only latent topics, but also employees’ “Expertise” and “Interests” from their microblogging activity. To measure employee \( i \)’s expertise on topic \( t \), we aggregate the number of words \( n_{ijt} \) assigned to topic \( t \) and author-recipient pair \((i, j)\), resulting in \( \sum_j n_{ijt} \). We quantify the result to a discrete scale of 1 to 5. The larger the value, the more knowledgeable an employee on a topic. Expertise is therefore a relative measure representing ones proficiency on topic \( t \), compared to other employees. To measure employee \( i \)’s interest on topic \( t \), we use the \( \alpha \) smoothed normalization and obtain \( \sum_j (n_{ijt} + \alpha) / \sum_t \sum_j (n_{ijt} + \alpha) \). We use this as a relative measure of the degree of preference an employee has on topic \( t \) with respect to other topics.

4.3 rESONAtE Workflow

Static and dynamic modeling of various data sources enables identification of topics, experts, connections and other relevant information across the enterprise. Expressed and latent information extracted in this manner is stored in rESONAtE. Discovered, new knowledge is provided to users in the form of recommendation of events, experts, connections, relevant content, etc. However, considering the dynamism in enterprise, the modeling should be updated regularly to keep track of temporal changes, as shown in Figure 9. Ontologies derived from static modeling and SNA tasks defined in dynamic modeling are utilized to define SNA task specification in Runtime, allowing execution of SNA tasks (LDA, ART etc.) at predefined intervals.

5 CASE STUDY ON REAL CORPORATE MICROBLOGGING DATA

In this section, we present a case study on a large scale dataset from a Fortune 500 company. The dataset is a complete snapshot of the in-house, corporate microblogging service. The functionality of the microblogging service resembles that of Twitter, whereas its interface is similar to Facebook. The dataset includes 9,855 unique users, and 15,200 messages with explicit reply links to other messages, over a time span of 1.5 years. The dataset contains users activity (message exchanges) and interactions (e.g. comment/reply, tagging), but lacks explicit social relationships (followee/follower) between users. Instead, we have acquired a snapshot of the organizational hierarchy with respect to users that participate in the microblogging service. Each employee is represented with a 4-character ID (e.g. “AECF”).
The organization level of each user is denoted using a character from “A” to “M”. Relationships are not symmetric: employee AECF may send a message to AAAF but AAAF can choose to not reply back. Further, AECF may be the supervisor of AAAF. Thus we have a directed labeled graph, with multiple relations between users (i.e. multiple Connection_Instance instances). Apart from structural closeness (i.e. being connected in the social graph), users may share common interests (see Sect. 4.2). We mine knowledge from message interactions between employees using the ART model (see Sect. 4.1).

5.1 Organizational Hierarchy vs. Informal Interactions

Organizational hierarchy is static, and dated, whereas communication may reflect “shortcuts”, i.e. collaboration that spans hierarchical levels. Seeking help, offering guidance, acting as mentor, thought leaders, influencer etc. study of communication may reveal change in organizational dynamics. For instance, employees Employee_AABG and Employee_AACD may appear as team-members according to organizational hierarchy, but reality (i.e. real life informal interactions between them in the corporate microblogging service) may hint otherwise.

Our intention is to better understand how information propagation works between corporate borders, and identify potential shortcuts in the organizational chart, as well as better understand how employees really collaborate to tackle everyday problems and provide solutions. To this end, we may examine employees’ connections under organization hierarchy connection context and compare these to “connections” due to interactions in the microblogging service.
Query 1: Given employee, find his connections with respect to organizational hierarchy.

```sparql
PREFIX resonate:<http://local.virt/DAV/home/dav/rdf_sink/user.rdf#>
WHERE {resonate:Employee_AABF resonate:hasConnection ?Connection.
  ?Position resonate:hasOrganizationLevel ?Organization Level.}
```

<table>
<thead>
<tr>
<th>Employee</th>
<th>Position</th>
<th>Organization</th>
<th>Level</th>
</tr>
</thead>
<tbody>
<tr>
<td>Employee_AAAD</td>
<td>Position_2</td>
<td>Organization_Level_D</td>
<td></td>
</tr>
<tr>
<td>Employee_AAAE</td>
<td>Position_3</td>
<td>Organization_Level_E</td>
<td></td>
</tr>
<tr>
<td>Employee_AAAF</td>
<td>Position_4</td>
<td>Organization_Level_A</td>
<td></td>
</tr>
</tbody>
</table>

Table 1. Partial result-set of a micro-analysis query listing connections of Employee_AABF, and their position in the Company.

5.2 Contextual Ego-Network & Community Identification

Processing and analyzing employee interactions reveals hidden dynamics. For instance, dynamic modeling that captures latent topics from discussions can be employed to reveal interest and expertise of specific users. At a micro level, topic oriented connections of a particular employee may reveal patterns that vary across topics. On the other hand, by clustering users with similar interest, it is possible to detect virtual communities around trending topics (macro level). A micro level query used to generate the topic gnostic ego-network of an employee follows.

Query 2: Given employee and topic, create a topic-sensitive ego-centric network.

```sparql
PREFIX resonate:<http://local.virt/DAV/home/dav/rdf_sink/user.rdf#>
SELECT ?User ?Topic
WHERE {resonate:Employee_AABF resonate:hasConnection ?Connection.
  ?Connection resonate:hasContentContext ?Interest.
  ?Connection resonate:hasconnectedTo ?User.}
```

<table>
<thead>
<tr>
<th>User</th>
<th>Topic</th>
</tr>
</thead>
<tbody>
<tr>
<td>Employee_AAAD</td>
<td>project</td>
</tr>
<tr>
<td>Employee_AAAD</td>
<td>connect</td>
</tr>
<tr>
<td>Employee_AAAD</td>
<td>pretty</td>
</tr>
<tr>
<td>Employee_AAAE</td>
<td>tag</td>
</tr>
</tbody>
</table>
5.3 Expert Identification

We mine employees’ expertise areas from their informal interactions. Their level of expertise may vary from topic to topic and from medium to medium. For instance, one might share innovative ideas and mostly contribute in discussions through emails, but not as much with respect to microblogging. We differentiate expertise according to communication channels (connection context), time and content (i.e. expertise varies with respect to email messages and microposts). A micro level query that retrieves topics and levels of expertise for a given employee follows. Expertise levels are quantized, taking integer values between 1 (less expert) and 5 (authority).

**Query 3**: Given employee, retrieve her areas of expertise and expertise levels.

```sql
PREFIX resonate:<http://local.virt/DAV/home/dav/rdf_sink/user.rdf#>
SELECT ?Topic ?Name ?Level
WHERE {
  resonate:Employee_AABG resonate:hasExpertise ?Expertise.
  ?Topic resonate:hasTopicName ?Name.
  ?Expertise resonate:hasExpertiseLevel ?Level.
}
```

<table>
<thead>
<tr>
<th>Topic</th>
<th>Name</th>
<th>Level</th>
</tr>
</thead>
<tbody>
<tr>
<td>Topic_97</td>
<td>share</td>
<td>Expertise_Level_1</td>
</tr>
<tr>
<td>Topic_84</td>
<td>drive</td>
<td>Expertise_Level_1</td>
</tr>
<tr>
<td>Topic_9</td>
<td>test</td>
<td>Expertise_Level_1</td>
</tr>
</tbody>
</table>

Table 3. Partial result set listing Employee_AABG’ expertise areas and corresponding levels.

5.4 Trends Macro-Analysis

Discovery of trending topic is a typical application in social network analytics. Typically, trending topics are identified for a single communication channel for single network. Such analysis at macro level may not be sufficient for an enterprise. In enterprise context, such macro analysis can prove to be more useful by discovering, for instance, trending topics and trending users for each communication channel, department, and organizational position. Given a trending topic, the following query retrieves its most prominent keywords, in all contexts. This is direct result of performing dynamic modeling over the enriched corpus.
Query 4: Given topic, get probability distribution of its trending keywords.

```
PREFIX resonate:<http://local.virt/DAV/home/dav/rdf_sink/user.rdf#>
SELECT ?Keyword ?Probability
WHERE {
  resonate:Topic 97 resonate:hasKeyword ?Topic Keyword No.
}
```

<table>
<thead>
<tr>
<th>Keyword</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>share</td>
<td>0.24534</td>
</tr>
<tr>
<td>interest</td>
<td>0.15791</td>
</tr>
<tr>
<td>friend</td>
<td>0.04047</td>
</tr>
<tr>
<td>facebook</td>
<td>0.0396</td>
</tr>
<tr>
<td>story</td>
<td>0.02046</td>
</tr>
<tr>
<td>easy</td>
<td>0.01176</td>
</tr>
<tr>
<td>found</td>
<td>0.01002</td>
</tr>
<tr>
<td>quick</td>
<td>0.00915</td>
</tr>
<tr>
<td>intern</td>
<td>0.00872</td>
</tr>
<tr>
<td>earlier</td>
<td>0.00741</td>
</tr>
</tbody>
</table>

Table 4. Keywords and probability distribution for given topic.

6 CONCLUSION

In this paper we introduced the enterprise dynamism leading to big data challenges. As solution to this challenge, we proposed to address integration and knowledge management aspects. Combination of semantic web techniques and social network analytics applied to handle latent and expressed semantics in the enterprise. We argued that in a typical enterprise, knowledge is always expressed and utilized in some context. We identified various kinds of contexts and proposed multidimensional modeling, which covers both static and dynamic communication aspects. We proposed a knowledge management workflow that enables, among others, continuous discovery of trends, expertise and interests, both at the employee and the enterprise level. Finally, we presented some use cases that demonstrate how our approach can be useful in practical enterprise setting. Our initial experiments on a representative microblogging dataset indicated vast potential of semantic social network analysis in addressing big, multidimensional data challenges for the enterprise. Our future work will focus on including additional (i) communication channels (ii) social network analytics techniques and (iii) advanced visualization capabilities.

REFERENCES


