DRDC-RDDC-2014-P5

"C2 in Underdeveloped, Degraded and Denied Operational Environments"

Automated extraction and characterisation of social network data from unstructured sources – An ontology-based approach

Topics

Topic 3: Data, Information and Knowledge

Authors

Étienne Martineau and Régine Lecocq
Defence Research and Development Canada – Valcartier
2459, Bravoure Road
Québec (Québec) G3J 1X5 Canada

Point of Contact

Régine Lecocq
Defence Research and Development Canada – Valcartier
2459, Bravoure Road
Québec (Québec) G3J 1X5 Canada
Tél.: 418-844-4000 x.4124
Téléc.: 418-844-4538
Regine.Lecocq@drdc-rddc.gc.ca

Abstract

Automated extraction of social network related data is one objective of the applied research project on SNA in Counter-Insurgency context (SNAC) at DRDC Valcartier. Since the vast majority of the information resides in unstructured text documents, the prototype must be able to extract social network related data directly from them. For these tasks, the prototype leverages and refines existing services provided by the Intelligence Science & Technology Integration Platform (ISTIP) at DRDC Valcartier. These services rely on ontologies to perform document annotation and to semantically characterise the data to be persisted. Given a list of a priory known instances of entities like people, organizations, and events, the system constructs the social web that ties these entities together. To do so, on a continuous basis, documents are fed via a data source crawling service that scans existing databases and returns new documents. Then, using natural language processing services, the system scans these incoming documents extracts and persists in a graph database information about entities as well as their relations and their respective attributes. The system also provides basic and semantic filtering services as well as conversion to many formats.

Introduction

Since the last decade, the Canadian Forces (CF) and their allies have been increasingly engaged in missions with an elusive and changing adversary who operates in intricate cultural environments. In this non-traditional context for the CF, social network analysis (SNA) is well positioned to enable analysts to better understand the composition and the mechanics of such culturally-different social networks. SNA can help in revealing and understanding the networks composition, structures, characteristics, as well as connections with other social networks. In their 2011 report [1], the United State Defense Science Board Task Force on Defense Intelligence stressed the critical importance of SNA in support of an Intelligence, Surveillance, and Reconnaissance capability in a COIN context. However, many challenges are being faced when attempting to perform meaningful SNA on covert networks [2] for the intelligence.

First, data about covert networks are, by definition, difficult to obtain [3]. Information about those networks is well guarded and, in general, not directly accessible and denied. Consequently, intelligence analysts must build their situational awareness based on an overabundance of indirect, degraded information and sources. Secondly, much of the information about social networks lays unrevealed inside unstructured reports pertaining to the theatre of operation. Whereas the human being is quite gifted in identifying meaningful information from documents, efficiently and effectively extracting data automatically from texts to enhance its utility remains an active research domain. Finally large heterogeneous social network data must be easily accessible into various formats to leverage existing SNA tools like ORA, Pajek or iGraph to name a few.

This document explains how the SNAC project tries to address some of the issues mentioned above. First, the SNAC project and the services from the ISTIP platform used by the prototype are introduced. Second, ontology and text-based templates relied upon for the for data extraction (a priory knowledge) that are defined. This will set the basis to understand how the automatic extraction system and the

associate's services were developed. Finally, the conclusion will address results and issues before discussing future works.

The SNAC project

Recently, DRDC Valcartier instigated an applied research project on SNA in a Counter-Insurgency context (SNAC). The SNAC project aims at investigating SNA techniques, tools and methods that could enhance the Intelligence Analysis capability. For exploration purposes, an integrative SNA prototype is being developed in the same line of thought than other authors [4], i.e., exceeds the sole analysis of social networks as pictured in figure 1. Indeed, it rather supports a SNA capability where services and functionalities cope with:

- the identification of meaningful data for the specific intelligence requirement for information (RFI);
- the automated extraction and organization of data on a regular basis;
- the selection of sub-networks of interest to be analyzed;
- the analysis of social networks; and
- the sensemaking brought to light through the contextualization of the analyses results.

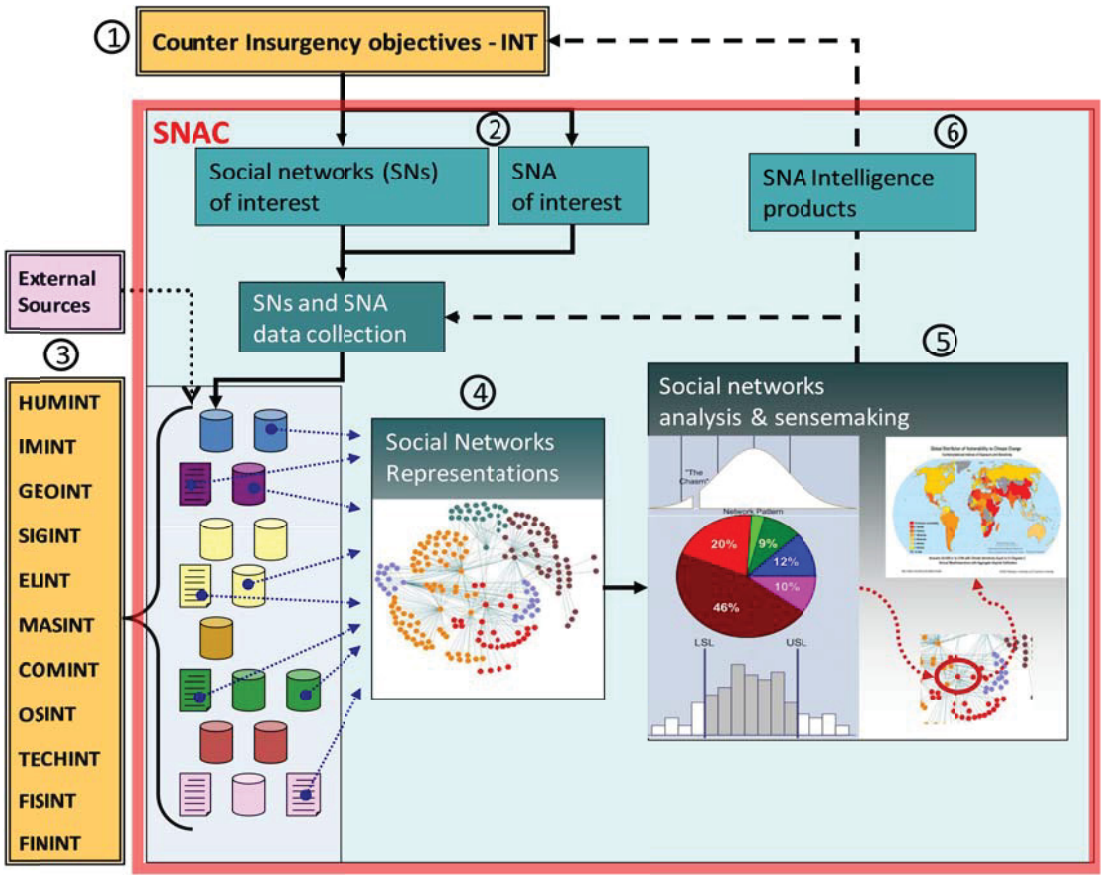

Figure 1: Approach for the SNA proof-of-concept prototype

The critical step to achieve this goal is to provide the analyst with meaningful data so that classical SNA method, statistics and visual analytics can be applied. Indeed, most commercial software provide extensive library of algorithm and analysis capabilities but provide no mean to gathers data from unstructured format. The SNAC project addresses this issue by providing an abstraction layer on top of external data sources to permit a structured view of the data. This functionality is deployed inside DRDC Valcartier Intelligence S&T Integration Platform (ISTIP) described in the next section. This layer is under the form of a standard web services interfaces on top of a graph database. The SNAC project data abstraction layer provides:

- single access point to up-to-date social network data;
- semantic, geospatial and temporal filtering;and
- different exportation format (CSV, GraphML, Pajek, etc.).

The ISTIP platform

The development of this SNA proof-of-concept prototype relies on and extends the services portfolio of the ISTIP platform. Following service-oriented architecture (SOA) design principles, the ISTIP provides a backbone platform for the iterative and incremental development of software inside DRDC Valcartier Command, Control and Intelligence (C2I) section. The ISTIP provides loosely coupled, reusable and composable services as foundations to intelligence support systems involved in different research projects. For example, the following capacity is offered on the platform:

- Automated reasoning of different kind (rule-based, case-based, description logic etc.);
- Anomaly detection;
- Visual analytics;
- Social network analysis;
- Ontology management;
- Natural language processing;

The problematic of data acquisition and extraction was identified at the outset of the SNAC research project and it was clear that an automated extraction capability would be a major enabler for a SNA capability. The project builds this capacity using services based on work done for the Multi-Intelligence Tool Suite (MITS) that had a similar capability [5]. Namely, the MITS is made of a module that extracts facts in an autonomous manner from unstructured text document. Since social network data can be expressed in the form of a fact, the SNAC project leverages the same services for automatic extraction of social network data which is:

- the Semantic Annotation of Text Documents (SATD);
- the Automatic Fact Extraction from Text Documents (AFEXTD);
- the Ontology Repository (OR); and
- the Document Repository (DR).

The two repository services mentioned above are quite straight forward. Like their name say they provide an access to ontologies and documents that reside on the ISTIP platform. They are an interface to a more complex data management system. The data can be accessed by File Transfers Protocol or directly by the file system. However, this service offers an easier access for computer system as well as notification capability when new data come in. The SATD service uses ontologies to annotate text documents. While this service main goal is to perform semantic annotation, it is also used to detect geolocation and time references as well as numerical values. The AFEXTD service extracts facts from unstructured text document based on patterns of words. These patterns are called the text-based template constraints and they are associated with a fact definition. When a document is processed, the text processing module first extracts the text from the document and then splits this text into multiple parts composed of one or many words within the sentence. In this case, the main types of aspects in which the text is split are:

- ontology entity occurrence (representing any ontology entity textually found in the text;
- lexical function of words (representing a lexical category function of a word or a group of word).

Each time a text-based template constraint is met inside a document, an associated fact is created. Details regarding automated fact extraction from unstructured text documents can be found in [6].

Ontologies and text-based templates

Ontologies and text-based templates compose the knowledge base that is used in order to search for social network data inside unstructured text document. Within the knowledge engineering community, a widely accepted definition of ontology was given by Gruber [7] as: “An ontology is an explicit specification of a conceptualization.” A conceptualization refers to an abstract model of some concepts, in ours case, social network and the COIN context. This definition has been extended by other authors to require that the conceptualization is shared (accepted by a group) and that the specification is not only explicit but also formally defined (machine-readable).

Ontologies are a central aspect of the SNAC proof-of-concept prototype and their usage go beyond the simple data accessibility first described in this document as stated in previous work [8]. For the abstraction layer on top of external data sources, ontologies are being used with respect to three essential aspects of data accessibility:

- the automated identification and extraction of social network related data;
- the persistence of the semantic registration of these social network data based on the context;and
- the creation of filters to prune only portions of the overall collected social network data.

As mentioned earlier, data extraction relies on existing web services of the ISTIP and on the knowledge base made of ontologies and text-based templates. The quality of the extracted data is dependant of the level of details of the ontologies. These are continuously enriched with as many concepts as possible in order to be able to cover all the key words used to describe social networks pertaining to the domain of

interest. To achieve this goal, a social network ontology was developed from a lexicon extracted from several years of operational documents and that looks like figure 2. Moreover, to be able to extract data from our allies, we also use a human terrain ontology developed by NATO. Examples presented in this document were prepared using Protégé-OWL. Web Ontology Language (OWL) is endorsed by the World Wide Web Consortium (W3C) to promote the Semantic Web Vision. OWL ontologies include classes, properties and instances. Relationships between classes are expressed as Object Properties (in contrast to data type properties which represent class attributes).

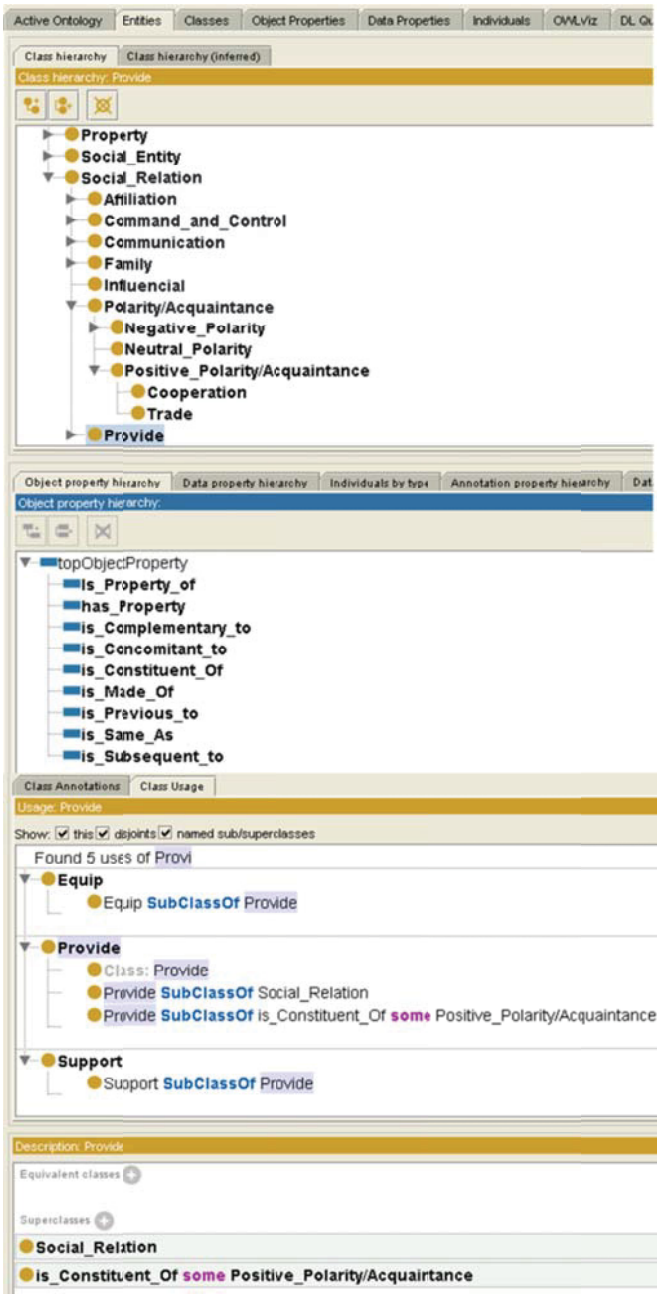

*Figure 2: S*ocial network ontology

Text-based templates define the patterns to look for inside a text document. Their use is similar to conventional regular expression but with an added semantic capability. Templates can refer to classes from ontology enabling the possibility to create more generic templates. Figure 3 shows a generic and a specific templates returning the same result if bind to an appropriate ontology, i.e., if the word "phoned" is an instance of the concept "communicate" (figures 3 and 4). Then again, the quality as well as the quantity of the extracted data depend on how well defined are the text-based templates.

As mentioned in the previous section, a text-based template is pair with a fact template. When the "constraint" is met, the fact is created. In the SNAC project these created facts for social networks have the underlying format of a triple (e.g.: subject-predicate-object as depicted in figure 4).

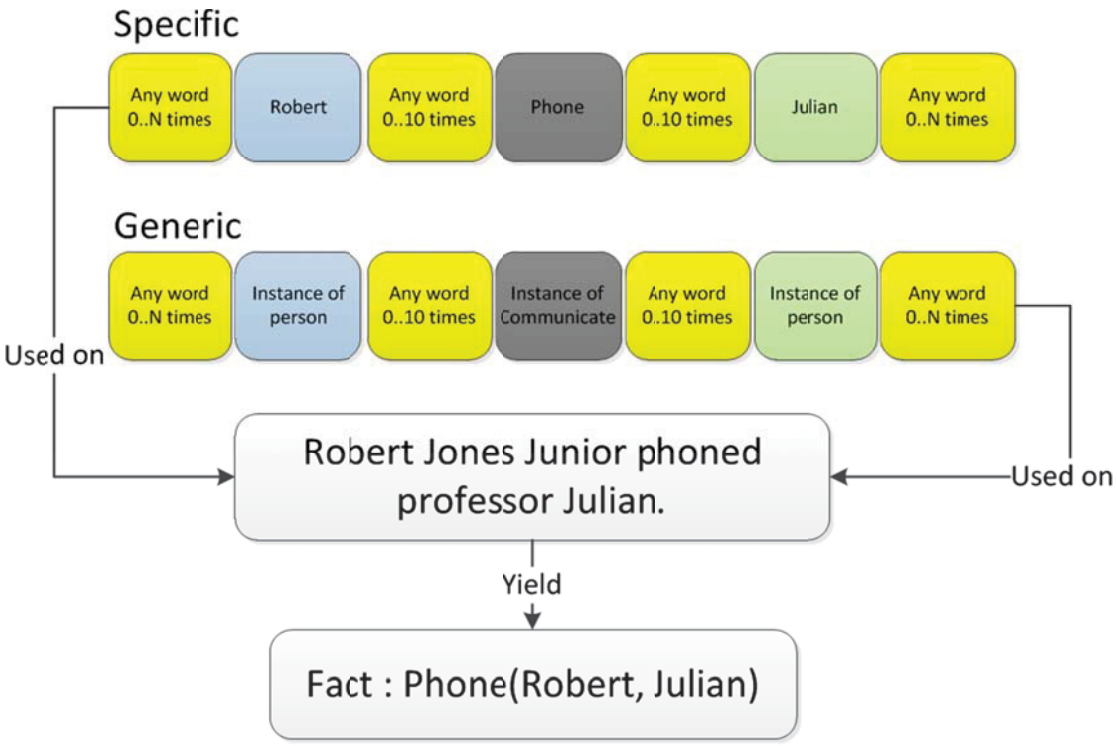

Figure 3: Comparison of generic and specific text-based template

Social network data extraction and persistence

In order to provide data through the abstraction layer, the incoming unstructured text document must first be processed to extract social network data. Then it must be organised in a meaningful manner so that an analyst can search inside it and increase his/her situational awareness. This step does not create new information, it just augments the value of the existing information by carving out the noise and by structuring the results.

To achieve this, the SNAC proof-of-concept prototype proceeds through various steps. The prototype is always waiting for new document notifications from the DR service. When a notification is received, the document is downloaded. The prototype also downloads the latest ontologies from the OR service since these can be modified or augmented at any time. The document and ontologies of interest are then sent

to the SATD service to create an annotated document. The result is redirected to the AFEXTD service along with text-based templates to create the facts about the social network. Since the facts are triples composed of instances of ontologies as depicted in figure 4, they can be converted into graph elements and persisted inside a graph database. The ISTIP platform possesses a Neo4J graph database [9] used for this purpose. However, the database does not always hold directly the annotated text, it rather holds references to instances inside ontologies. The exception is for numerical attributes like data, age, and geolocalisation.

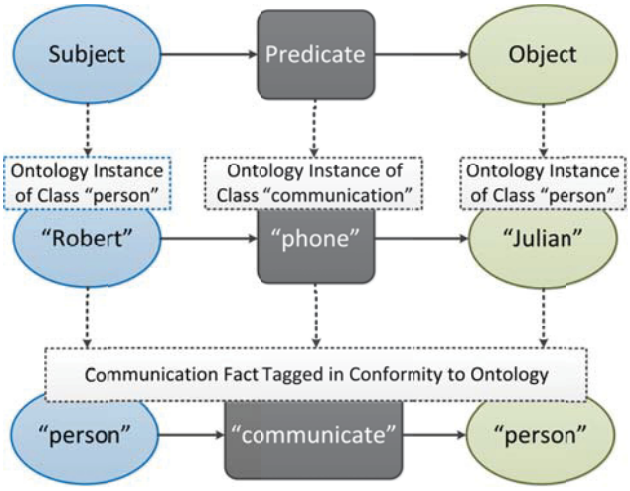

Figure 4: Example of a fact following the triple format.

The nodes, links and attributes returned by the extraction (called semantic graph model) are in a format that cannot be saved directly inside the native format of the Neo4j database (graph data model). The challenge faced here is that the Neo4j format cannot store attributes on attributes, only on nodes and relations. This is a problem as such attributes are a mandatory requirement for our military client. To solve this problem it was decided to store relations inside Neo4j nodes. One of the most important aspects of this design is that it represents an Entity's properties and relations using nodes, instead of representing properties as the entity's node's attributes and relations as edges in the graph. The side effect of this design is that it requires a much higher number of nodes in the graph and makes data in the graph much less intuitive. On the other hand, it allows properties and relations to be indexed using an in-graph index and therefore supports efficient graph traversal. Also, it provides the flexibility to put properties (like timestamp) anywhere on the graph. However, this makes direct requests to the database complicated since they must take into account the encapsulation of relations and attributes, i.e., convert requests to the graph data model and convert results to the semantic graph model. Figure 5 pictures a graph data model view of the same data as in figure 6 which is a semantic graph model view. To handle this challenge, the prototype provides a data access service that performs all the required conversions.

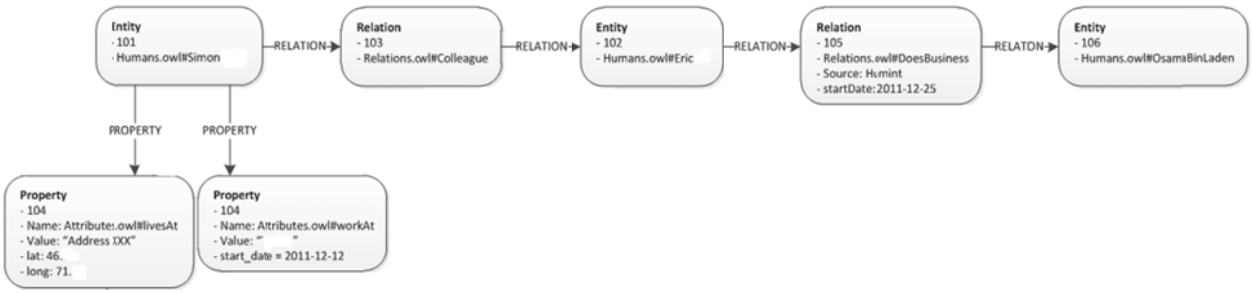

Figure 5: Graph data model view example.

Figure 6: Semantic graph model view example.

Social network data access services

Once the data is persisted in the graph database, it must be made available to other functionalities through the ISTIP platform. To meet the need of the intelligence request for information, the social network data abstraction layers must be able to provide different views of the data. This is critical as first, the data gathered by the automatic extraction can provide large heterogeneous networks that must be filtered before any type of analysis can be performed over them. Second, the analyst may be using different tools requiring different input formats to analyse the data. Finally, as mentioned in the previous section, the data inside the Noe4j database is not directly usable as it stands and need to be converted. Therefore, the abstraction layer must be able to filter the data and return it under various formats. To achieve this, the prototype instantiates three new services:

- the filter characterisation service;
- the network filtering service; and
- the social network data conversion service.

To be able to perform basic filtering as well as semantic filtering on the data, a user must invoke the filter characterisation service. This service is needed to convert a semantic query to a format understandable by the Neo4j database of the ISTIP platform (going from the semantic graph model as depicted in figure 6 to the graph data model from figure 5). This service uses domain ontologies to identify all possible instances in the ontologies that a semantic query may be referring to (nodes or relations). This characterisation can include temporal references to retrieve social entities inside a time interval like dates of events or a time window, time of reporting or time of database insertion. It is also possible to specify a region of interest by name, near a point of interest or inside a polygon to filter geotagged social entities (Fig 7). Moreover, logical operation can be applied to the characterisation

(AND, OR, NOT, XOR) as well as numerical comparator for attribute (inside interval, greater than, lower than, etc.). The resulting query is then passed to the network filtering service that extracts the corresponding data and returns it in the graph data model format.

The resulting data needs to be converted since it is not intelligible in the graph data model format. Inside the SNAC prototype, data is handled under the semantic graph model format. However, different library and software currently in use do not support this format and put specific restrictions on how the data should be provided to them. While most external libraries/systems/softwares support CSV and GraphML, to provide the maximum flexibility, the social network data conversion service offers many more output formats

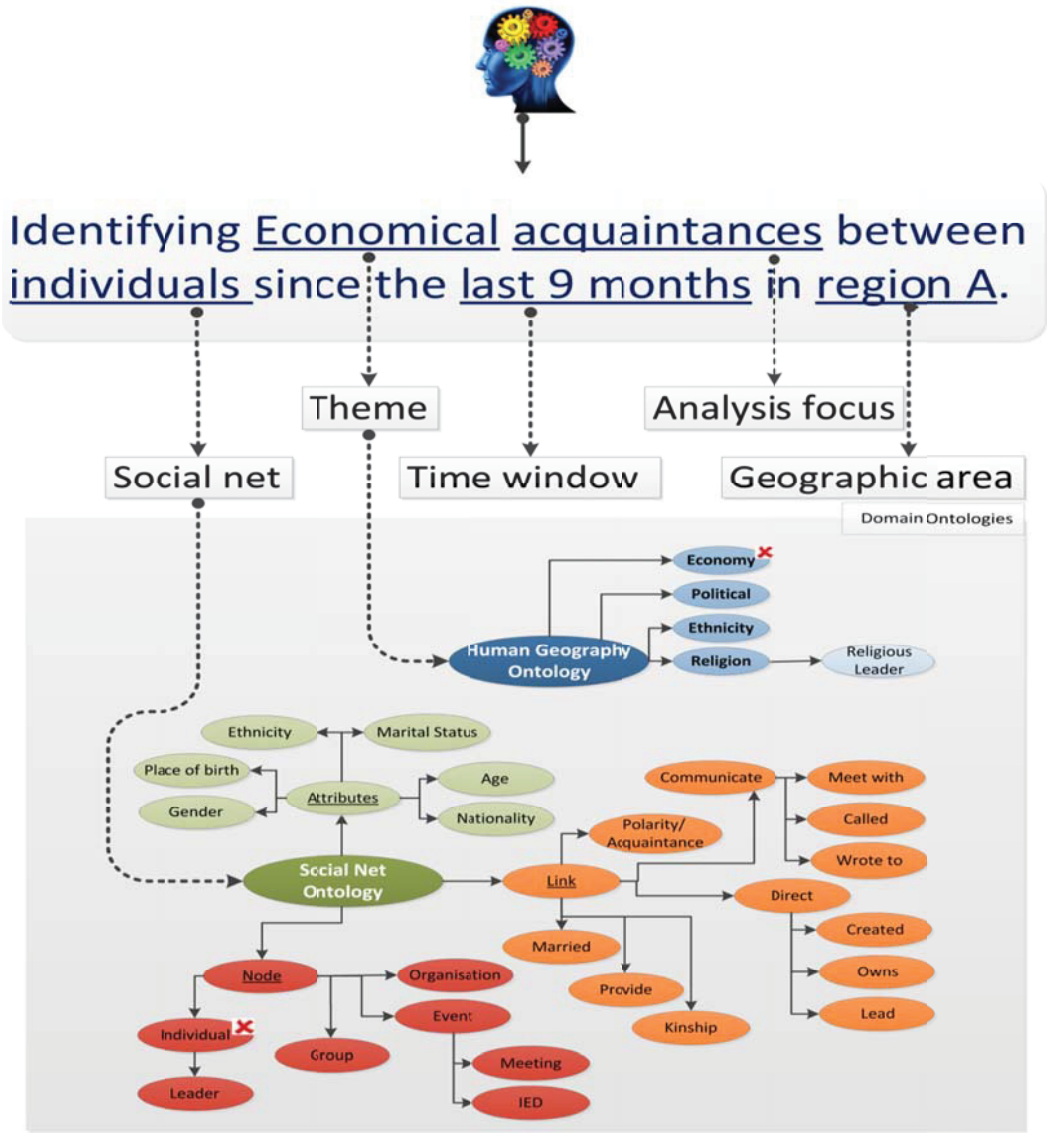

Figure 7: Query characterisation.

Conclusion and discussion

The SNAC research project aims at providing a full SNA capability. At the present time, automated social network data extraction systems remain an area of intensive research. In this regard, the approach considered through the SNA proof-of-concept prototype has proven to be a good solution so far. It provides an easy access to social data scattered inside large quantity of unstructured text document. For the analyst, it gives the possibility to visualize (using link analysis means) and perform analyses on up-to-date social network data. Moreover the filtering capability can extract sub-networks of interest based on semantic, time geo-localisation and properties as well as some basic logical operations on sets. The data can be used inside the ISTIP or with third party tools like iGraph library or some commercially of the shelves softwares by using the conversion service that can output to formats like GraphMl, CSV and Pajek.

However, the automated extraction system of the SNAC project still faces some challenges. First, the ontologies must be augmented with known entities. Currently, if the system encounters an entity that is not inside the ontology, it will be ignored and no data will be extracted about it. The system only completes the a priori knowledge of the analyst about known entities, it cannot discover new ones. Second, it is difficult, even with the right ontology to create text-based templates that will capture all the information. Misspelling, aliases, abbreviations or unconventional writing styles are not handled by the system. Finally, these problems are hard to quantify since the only way to validate the results is to go manually through all the documents. The system extracts a lot of data but it is difficult to evaluate its efficiency in a rigorous way.

Future work

The future work on this aspect of the project will consist in building more complete ontologies and text-based templates to extract more information. There is also an ongoing validation procedure that compares the extracted information about the social network by the analyst against the ones extracted by system. Also, a comparative evaluation of the performance of the analyst with or without the prototype will be conducted on historical operational data from the Canadian forces. In the meantime, tests with unclassified data are also planned. To conclude, the automated extraction is a stepping stone to achieve a social network analysis capability. The quality and quantity of information extracted shows great promises for the following analyses steps of the prototype.

References

[1] USD (AT&L) - Office of the Under Secretary of Defense (Acquisition Technology and Logistics), “Report of the Defense Science Board Task Force on Defense Intelligence - Counterinsurgency (COIN) Intelligence, Surveillance, and Reconnaissance (ISR) Operations”, Washington DC, 2011.

[2] É. Martineau and R. Lecocq, “Analysis of Social Networks in COIN Context”, presented at the IST-112 Symposium on “IST-112/SET-183 Joint Symposium on Persistent Surveillance: Networks, Sensors, Architecture, Quebec, Canada, 2012.

[3] Roberts, N. and Everton, S.F., “Strategies for Combating Dark Networks“, Journal of Social Structure, vol. 12, pp. 1-32, 2011.

[4] A. Semenov, J. Veijalainen, and A. Boukhanovsky, A., “A Generic Architecture for a Social Network Monitoring and Analysis System” Network-Based Information Systems (NBiS), 14th International Conference on , vol., no., pp.178-185, 7-9 Sept. 2011.

[5] Roy, J. and Auger, A., The Multi-Intelligence Tools Suite – Supporting Research and Development in Information and Knowledge Exploitation, Proceedings of the 16th International Command and Control Research and Technology Symposium (ICCRTS), Québec City, Canada, 21–23 June, 2011.

[6], Auger, A., *Acquisition and Exploitation of Knowledge for Defence and Security*, NATO IST-087 Symposium on Information Management / Exploitation, Stockhlom, Sweden, 19-20 October 2009.

[7] Gruber TR “A translation approach to portable ontology specification.” Knowledge Acquisition 5(2): 199-220. 1993.

[8] Lecocq R., Martineau E., Caropreso M. F., An Ontology-based Social Network Analysis Prototype, Proceedings of the 3rd IEEE Conference on Cognitive Methods in Situation Awareness and Decision Support (CogSIMA 2013), San Diego, CA, USA, 26–28 February, 2013.

[9] The Neo4j, NOSQL graph database. DOI=http://neo4j.org/