Decision Support in Open Source Intelligence Applications

Ortiz-Arroyo, Daniel

Published in:
Studies in Computational Intelligence

DOI (link to publication from Publisher):
10.1007/978-3-319-08624-8_5

Publication date:
2014

Decision Support in Open Source Intelligence

Daniel Ortiz-Arroyo

Computational Intelligence and Security Laboratory
Department of Electronic Systems
Aalborg University
do@es.aau.dk

Summary. This chapter describes a decision support system (DSS) specially designed for open source intelligence. The decision support system was developed within the framework of the VIRTUOSO [1] project.

Firstly, we describe the overall scope and architecture of the VIRTUOSO platform. Secondly, we describe with detail some of most representative modules of the DSS. The modules employ intelligent computing techniques such as knowledge representation, soft-fusion and fuzzy logic.

The DSS together with other tools developed for the VIRTUOSO platform will help intelligence analysts to integrate diverse sources of information, visualize them and have access to the knowledge extracted from these sources.

1 Introduction

Decision support systems (DSS) comprise a set of computational tools whose purpose is to support better decision making processes within organizations.

DSS allow decision makers to visualize, analyze, process and mine data of various types. Data is collected from a diversity of data sources and integrated to create a knowledge based repository. The knowledge base is used by decision makers and analysts to help them in reasoning about some possible scenarios.

DSS are used in a variety of fields such as business intelligence and criminal intelligence. In the area of business intelligence, DSS help companies to asses competitors market position, the market trends and plan future investments. In criminal intelligence, DSS help intelligence agencies to tackle organized crime.

DSS have become more sophisticated in the internet era. The internet has eased the communication and sharing of information. Moreover, social media

[1] The Versatile InfoRmation Toolkit for end-User oriented Open-Sources explotatiOn (VIRTUOSO) project was developed between 2010-2013 with funding from the European commission within the seventh framework programme (FP7). Information about the VIRTUOSO project is available at http://www.virtuoso.eu

is increasingly used by hundreds of millions of people, including government and business organizations but also by organized crime.

Traditional mass media uses the internet for reporting news about countries, people, government, and events using different types of data sources such as RSS feeds, blogs and specialized news portals containing video, audio and text.

The internet has also eased the distribution of specialized technical information and computer code that can be used to exploit the weaknesses of IT systems. This has made easier for hackers and organized crime to carry out cyber-attacks against specific servers of institutions or to the whole computer network of a country. These attacks are carried out using a botnet of "zombies" computers distributed around the world [1].

Traditionally, the data sources used by intelligence analysts to track down organized crime were *classified*. Classified, secret data about individuals or organizations is kept in secure data repositories isolated from the internet.

However, intelligence agencies have recognized the value that the information publicly available on the web has in their investigations. For instance the usage of social media by organized crime and its associates may leave intended or unintended traces that can be collected but this must be done respecting the legislation associated with information gathering in a country[2].

Open information collected from the web is being used in a variety of areas such as criminal investigations, situation monitoring and assessment, and to produce early warnings of possible crisis.

The collection of methods used in collecting, managing and analyzing publicly available data is called *Open Source Intelligence* (OSINT).

Some of the data sources employed in OSINT are electronic media such as newspapers and magazines, web-based social media such as social networks, web pages and blogs, public data from government sources, professional and academic literature, geospatial data, scanned documents, video and data streams among the most common sources.

OSINT creates important technical, legal and ethical challenges. One of the main technical challenges is to collect relevant meaningful information from reliable sources, among the huge amount of data sources available on the web. This is a critical issue since information may be of low reliability or bogus, obsolete, duplicated and/or available only in certain languages. Information may be of different types and being available in different formats.

Once a reliable data source is found, the relevant information on it must be identified, extracted and stored in a knowledge base. The extraction of relevant information from documents is also technically challenging, since entities expressed in natural language such as events, individuals, organization and places must be recognized and disambiguated.

After the extraction process, information is stored in knowledge bases, a step that coverts raw information into knowledge. The knowledge stored in knowledge bases is commonly represented in the form of ontologies and semantic networks. These ontologies contain specialized and general domain knowledge about entities such as individuals, organizations and events.

Knowledge bases allows analysts to reason about the entities stored in them and their relationships.

DSS provide also visualization tools, allowing the analyst to look at the data from different perspectives. Some of the commonly used visualization tools include dashboards, graph viewers and editors, maps and plots of different types.

The technical aspects of an OSINT system are very important, but they are not the only challenge that must be addressed when designing and implementing an OSINT system.

The other challenge in OSINT are related to legal and ethical aspects. In OSINT relevant information must be collected in a way that respects the privacy of individuals and a the same time does not violate the existing national and international laws.

Collecting personal information about individuals and organizations from the numerous open (and proprietary) sources on the internet has opened up the possibility for using such data for mass surveillance purposes [2].

This issue has created a debate on the ethical and legal aspects involved in the use of OSINT-based technologies that must be addressed.

In the case of the VIRTUOSO project, these issues were addressed since the conception of the project. One of the tasks continuously performed during the whole development process of VIRTUOSO, was to make sure that privacy was respected and that no laws (for instance copyright) were violated when collecting and storing data.

The VIRTUOSO platform addressed all the technical and ethical challenges described above.

The DSS included in VIRTUOSO employs intelligent computing techniques that allows the OSINT system to manage uncertainty, represent and fuse knowledge, and process natural language among other tasks.

This chapter is organized as follows. Section 2 describes the architecture of the VIRTUOSO platform. Section 3 describes some representative modules of the decision support system in VIRTUOSO. Finally, section 4 presents some conclusions.

2 VIRTUOSO's Architecture

In summary, the goal of the VIRTUOSO project is to retrieving unstructured data from open sources available on the web and converting it automatically into structured actionable knowledge. To achieve this goal a flexible architecture for the whole system was designed.

[2] The recent revelations by E. Snowden about the mass surveillance programs deployed by US intelligence agencies has revived the debate within Europe and in other parts of the world about this issue

The architecture of VIRTUOSO is based on a Service Oriented Architecture (SOA).

SOA was proposed to ease communicating, synchronizing and integrating diverse software components that implement services. This was an extremely important issue in VIRTUOSO, because software components could be developed by different teams, using different languages and technologies.

SOA is not an implementation but a recommendation for how to structure component systems based on services.

In VIRTUOSO the SOA model was implemented using the Weblab [3]platform. Weblab is an open source intelligence platform, whose main purpose is to build software systems specifically for OSINT applications based on the SOA specification.

The architecture of VIRTUOSO consists of three main processing stages: a) data acquisition, b) data processing, and c) decision support.

In the data acquisition stage, data from unstructured open sources is retrieved using web crawling techniques. Web crawlers acquire different types of data from a wide diversity of sites on the web i.e. electronic-text data, multimedia content, and even from scanned papers. These multiple types of data come from web sites, blogs, tweets, RSS feeds, trends, video streaming sites, and paper documents.

Regarding electronic-text data, at the current state of the project more than 500,000 documents are processed every day, written in 39 different languages from 188 countries. These documents are retrieved from 28,000 open sources.

The data acquisition stage is continuously connected to the Internet to retrieve all relevant data. Additionally, at this stage some pre-processing is performed. For instance, normalization, object recognition, entity naming, event extraction, image and video classification, source assessment, and speech recognition.

Normalization of different types of media and documents is performed by representing them in a single XML based format that contains pointers to the real location of data. The source assessment stage attempts to evaluate the reliability of a data source.

The number of pre-processing steps performed by VIRTUOSO platform can be configured.

After the pre-processing stage, a special data repository is created, containing all the results of all pre-processing steps that may have been performed. Figure 2 shows a high level view of the data acquisition and pre-procesing stages of VIRTUOSO.

Both, the data acquisition and preprocessing stages were implemented by integrating all its components on the SOA model.

Contrarily, the data processing stage of VIRTUOSO is not connected directly to the internet. This is mainly done for security reasons.

[3] Weblab is available at weblab-project.org

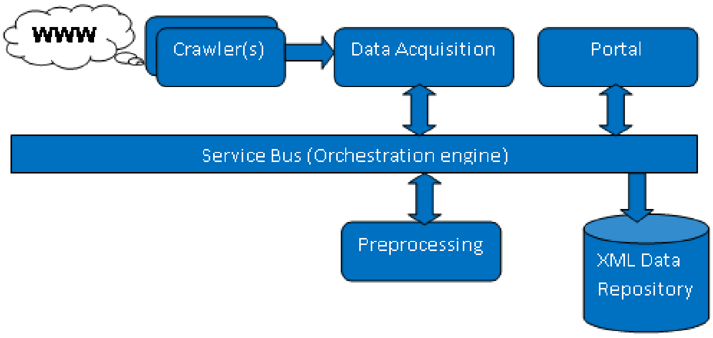

Fig. 1. Data acquisition, preprocessing and data repository

The data processing stage contains several components, among which are: a full text and multimedia search engine, a summarization component, automatic translation of documents, determination of document similarity, and query translation.

The knowledge base is one of the key components in VIRTUOSO. The knowledge base is created apriori with general domain knowledge and is updated with knowledge extracted in the data pre-processing stage.

To being able to use the data repository created in the pre-processing stage, an import/export module is available at the processing stage. During importing, data may be manually or semi-manually validated to ensure that no irrelevant or dubious data is introduced in the knowledge base.

The knowledge base is part of the decision support system contained in VIRTUOSO.

The data processing stage and the decision support system, share the same SOA-based Weblab infrastructure.

Figure 2 shows a high level view of the data processing and decision support modules of VIRTUOSO.

The decision support system is described in the following section.

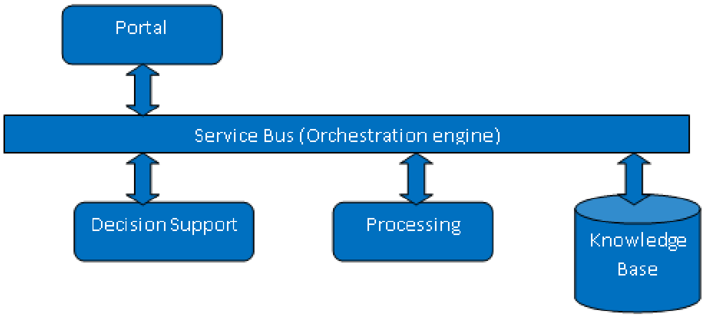

Fig. 2. Data processing, knowledge base and decision support system

3 The Decision Support System

The decision support system available in VIRTUOSO is one of its key components.

VIRTUOSO's decision support system consists of several software modules that can be applied to process documents, directly to data or to both data and documents.

The modules that can be applied to documents are: metadata viewer, source assessment, geographical search, multimedia and text search, trend analysis, social media topic, sentiment monitor and semantic search.

The modules that can be applied to data are: graph viewer, tabular viewer, graphical SPARQL querying, rule editor, entity editor, similarity of entities, similarity of strings, semantic analysis, and social network analysis module.

The modules that can be applied to both data and documents are: dashboard visualization, knowledge browser, centipede and traceability module.

These modules allow the analyst to perform a wide variety of tasks from querying semantic knowledge bases, to the visualization and processing of different types of graphs, data and documents. Some software modules are not used directly by the analyst, but instead they provide services to other modules in the system.

All the modules included as part of the DSS work seamlessly together on the Weblab SOA platform.

Given that is not possible to describe in detail all the modules included in the DSS of VIRTUOSO, in the rest of the chapter we describe a few of the most representative modules. The documentation about the DSS in VIRTUOSO contains a detailed description of the rest of the modules [3].

3.1 Knowledge Base, Fusion and Uncertainty Management

The knowledge base in VIRTUOSO comprises ontological knowledge (conceptual and geographic) and operational knowledge (factual).

The ontological knowledge consists of the known existing relationships among the concepts employed in the intelligence domain. The knowledge base contains several ontologies that include knowledge about general and specialized domains.

Ontological knowledge is represented as triples in the form (predicate, subject, object) or (p,s,o) for short, as specified in the resource description framework (RDF) schema. Internally the knowledge base employs a slightly different format based on RDF.

Semantic knowledge in VIRTUOSO may be introduced in the knowledge base either manually for highly specialized domains or in an automatic or semiautomatic way for other types of domains. For instance, part of the ontological knowledge included in VIRTUOSO is imported from existing ontology resources. However, to use these or other existing ontology resources a process of semantic disambiguation and fusion of information is performed.

The fusion module in VIRTUOSO merges two graph structures, using an operation called “maximal joint”. This method was originally proposed to fuse conceptual graphs in [4]. However, in VIRTUOSO the maximal joint heuristic was applied to graphs.

The joint operation was divided into two parts. First, the compatibility of the two elements to fuse is evaluated. Two entity nodes are considered compatible if the type of entity is the same (e.g person, location) and if a high proportion of entities’ properties is similar. The similarity measure applied depends on the type of properties that entities may have.

In VIRTUOSO the nodes in the graph structures correspond to entities that have properties defined as strings of characters or numbers.

The similarity of string properties, like names for instance, was evaluated using Levensthein string edit distance. This distance basically evaluates how many insertions or deletions of characters are needed to convert one string into the other.

For numerical properties, the similarity was calculated using different techniques. For instance, in the case of date properties the number of days between the two dates was used as the distance. For other types of numerical properties the similarity was evaluated using the following equation:

$$sim_{num}(\beta, x, y) = e^{\frac{\beta(x-y)^2}{\beta-1}} \tag{1}$$

where β represents the sensibility of the measure to the distance between two similar numerical values x and y.

Two entities were considered compatible if the similarity value of the numerical and string properties was above certain threshold value.

The maximal join operation was used to fused the sub graphs of two distinct but compatible graphs using the following method. Two compatible nodes were fused, creating an extended graph that included the sub-graphs of the two compatible concepts. This procedure was repeated recursively in each node of the subgraphs until incompatibilities were found.

The operational knowledge in the knowledge base consists of information extracted from the open sources. The basic entities available in the knowledge base are physical entities (e.g., persons, vehicles), legal entities (e.g., organizations), non-physical entities (e.g., phone number), and event entities (meeting, travel).

The knowledge base contains also various types of metadata, such as time dependencies, validity (or certainty), sensitivity, confidentiality, and provenance information (to being able to trace back to the sources).

To mange uncertainty the RDF triples in the knowledge base were extended by adding a parameter β that depending on the entity and type of uncertainty may represent a probability distribution or a possibility distribution. RDF triplets were then represented as $\{(\text{p,s,o}),\beta\}$ using that information.

3.2 Social Network Analysis Module

Social Network Analysis (SNA) comprises the study of relations, ties, patterns of communication and behavioral performance within social groups. In SNA, a social network is commonly modeled by a graph composed of nodes and edges. The nodes in the graph represent social actors and the links the relationship or ties between them [5].

Being criminal organizations also a form of social network, they are represented as graphs in the user interface of VIRTUOSO. The nodes in the graph represent the members of an organization and the links represent their known relationships. In general, the relationships may be known connections between individuals or may represent the structure of command within an organization. These connections may be manually introduced or extracted from a document collection and stored in the knowledge base.

In SNA, multiple metrics have been proposed with the aim of evaluating the importance of the nodes within a social network. One of the most important metrics in SNA is *centrality* [6][7]. Centrality describes a member's relative position or importance within the context of his/her social network.

One of the applications of the centrality measures used in SNA is to discover *key players* [8]. Key players are these nodes in the network that are considered "important" in regard to some criteria, such as number its connections, their role in diffusing information, and their influence on the network, among other criteria.

To process the social networks, the SNA module employs some of the most popular centrality measures used in SNA such as degree, betweenness, closeness, and eigenvector centrality [9].

The analyst should decide which centrality measure is the most appropriate to apply in a particular situation. However, in certain cases the analyst may be interested in evaluating the overall importance of a group of nodes. In this case, it is possible to calculate an aggregated centrality value using all of the centrality measures available in the SNA module. The aggregation could be also useful when the analyst is not sure about which centrality measure should be used.

To perform this calculation, the SNA module in VIRTUOSO employs an ordered weighted aggregation (OWA) operator [10]. The OWA operator is defined as:

$$h_w(a_1, a_2, ..., a_n) = w_1 b_1 + w_2 b_2, ... + w_n b_n \tag{2}$$

where $w_i \in [0, 1]$ are the weights and $\sum_{i=1}^{n} w_i = 1, (b_1, b_2, ..., b_n)$ is a permutation of vector $(a_1, a_2, ..., a_n)$ in which the elements are ordered $b_i \geq b_j$ if $i < j$ for any i, j.

The weights used in the OWA operator are normally calculated using its *andness* value. The *andness* value of the OWA operator is defined as:

$$Andness(\mathbf{w}) = 1 - \alpha = 1 - \frac{1}{n-1}\sum_{i=1}^{n} w_i(n-i), \alpha \in [0,1] \tag{3}$$

where α is the *orness* value. This *andness* value represents how close the OWA operator behaves as a fuzzy *and* operator i.e. how close the resulting aggregation value is to the fuzzy AND (minimum) value produced by all the centrality measures. For instance, with *andness* value of 1 the weights of the OWA will be (1,0,...,0), which will produce the minimum value in the aggregation. With an *andness* value of 0.5 all weights will be $1/n$ and the OWA will calculate the average value of all its inputs.

When OWA operators are used, expert knowledge about a problem domain is used to decide the most appropriate *andness* value. In the prototype of the SNA module weight values of (0.1,0.15,0.25,0.5) were assigned by default to the OWA operator. These weight produced an *andness* value of 0.71. Hence, the default centrality measure produced by the OWA operator will be a value that lies between the average centrality of all the measures and the minimum value produced by each of them.

One of the issues found when using an OWA operator, is that very different values in the weights may produce the same *andness* value. Therefore, one desirable feature of an OWA operator is to get good dispersion in the weight values, in a way that most inputs contribute to produce the desired *andness*. This issue has been addressed by other aggregation operators, such as maximum entropy OWA (MEOWA) operator [11] and the andness-directed weighted aggregation (AMWA) operator [12]. All these operators have been also implemented in the SNA module in VIRTUOSO.

It must be noted that for some specific constant values of *andness*, it is possible to use some of the analytical expressions described in [13] to calculate the weights used in the OWA operator. As is described in [13] these expressions provide good dispersion values in the weights' distribution. For instance for an *andness*=2/3=0.66 we could use the following equation to calculate each of the n weights used in the OWA operator:

$$w_i = \frac{2i}{n(n+1)} \tag{4}$$

using this expression, the OWA weights will be $(2*1/(4*5) = 0.1, 2*2/20 = 0.2, 2*3/20 = 0.3, 2*4/20 = 0.4) = (0.1, 0.2, 0.3, 0.4)$, which have good weight dispersion and whose values are close to the values produced by the *andness* of 0.71 used by default in the SNA module.

The application of the OWA, MEOWA or AMWA operators allows the analyst to use all the centrality measures available at once, in such a way that each one contributes partially to the overall calculation of the centrality of every node in the network.

The SNA module can be used used in two different ways in VIRTUOSO. One way is as a REST web service that receives HTTP/POST requests containing the description of the social network to be analyzed. This description is provided in a standard format such as graphml [4]. The output of the service is the calculation of the desired centralities for each node in the network. These values are returned in JSON format encoding.

It is also possible to use the SNA module functionality integrated as a portlet [5] within the weblab platform.

4 Conclusion

VIRTUOSO provides a large collection of software components and modules that help analysts to process and visualize a large collection of data of various types.

These modules together with the decision support system and the knowledge base containing extracted semantic information about entities, will help the analyst to reason more easily about a particular scenario.

VIRTUOSO is a complex system and we have described just a few of its features. At the current stage the decision support system and all the tools developed in VIRTUOSO have been tested on a few scenarios and a final presentation on the results of the project has been performed. The software modules available in VIRTUOSO are at pre-release state at the time of writing this chapter.

One of the challenges in using complex software systems like VIRTUOSO, is how to used them in the most effective way. VIRTUOSO requires a group of IT specialists, administrators and analysts that could manage the data sources, maintain the knowledge base, define the most relevant scenarios, and analyze the results provided by the system.

The final report on the VIRTUOSO project included some recommendations in this regard. However, the procedures employed and the type organization that may be needed to use effectively the whole system must be tailored to fit each specific organization.

As part of the project, a demonstrator of the VIRTUOSO platform will be created and installed at the Laboratory of Computational Intelligence and Security at Aalborg University campus Esbjerg in Denmark.

Interested parties within the European Union will be allowed to use the demonstrator and experiment with the system, to asses its functionality.

5 Acknowledgments

The author was member of the team from Aalborg University that participated in VIRTUOSO. The research leading to these results has received funding from the European Union's Seventh Framework Programme (FP7/2007-2013) under grant agreement no. 242352.

[4] graphml is an XML-based format used to represent graphs or networks

[5] portlets are pluggable user interface components that are managed by web portals

References

1. Ortiz-Arroyo D. Information security threats and policies in europe. In Laundon K. and Laundon J., editors, *Management Information Systems: Managing the digital firm. Global Edition: Managing the Digital Firm*, pages 357–358. Pearson Longman, 2011.
2. R. Frank, C. Cheng, and V. Pun. Social media sites: New fora for criminal, communication, and investigation opportunities. Technical Report 21, Simon Fraser University, 2011.
3. Virtuoso Consortium. Decision support and visualization modules report. Technical report, Virtuoso Project Consortium, 2013.
4. C. Laudy, E. Deparis, G. Lortal, and J. Mattioli. Multi-granular fusion for social data analysis for a decision and intelligence application. In *Proceedings of 16th International Conference on Information Fusion*. IEEE, 2013.
5. J. Scott. *Social Network Analysis: A Handbook*. SAGE, 2000.
6. L. C. Freeman. A set of measures of centrality based on betweenness. *Sociometry*, 40(1):35–41, 1977.
7. N. E. Friedkin. Theoretical foundations for centrality measures. *The American Journal of Sociology*, 96(6):1478–1504, 1991.
8. Ortiz-Arroyo D. Discovering sets of key players in social networks. In Abraham A., Hassanien A., and Snasel V., editors, *Computational Social Networks Analysis: Trends, Tools and Research Advances*, pages 27–47. Springer, 2010.
9. S. P. Borgatti and M.G. Everett. A graph-theoretic framework for classifying centrality measures. *Social Networks*, 28(4):466–484, 2006.
10. R. R. Yager. On ordered weighted averaging aggregation operators in multicriteria decision making. *IEEE Transactions on Systems, Man and Cybernetics*, 18:183–190, 1988.
11. D. Filev and R. R. Yager. Analytic properties of maximum entropy owa operators. *Information Sciences*, 85:11–27, 1995.
12. L. H. Larsen. Multiplicative and implicative importance weighted averaging aggregation operators with accurate andness direction. In *Proceedings of the Joint 2009 International Fuzzy Systems Association World Congress and 2009 European Society of Fuzzy Logic and Technology Conference*, pages 402–407. IEEE, 2009.
13. B. S. Ahn. On the properties of owa operator weights functions with constant level of orness. *IEEE Transactions of Fuzzy Systems*, 14(4):511–515, 2006.