Hidden information retrieval and evaluation method and tools utilising ontology reasoning applied for financial fraud analysis.

Mariusz Chmielewski[1,a], *Piotr* Stąpor[1,b]

[1]Military University of Technology, Cybernetics Faculty, gen. W. Urbanowicz Street 2, Warsaw, Poland

Abstract. The paper summarizes a semantic association evaluation and reasoning method, utilising domain and problem solving ontologies. The method combines algorithms for data aggregation and logic reasoning utilising concrete financial data. As an outcome method supports suspicious behaviour recognition of money loundering schemes. These scenarios and schemes are implemented by the analysts using ontology-based constructs. Provided tools cover all stages of data processing starting from structural data extraction and migration, aggregation to reasoning using logic (DL and FOL) constructs. Advances in automatic reasoning and the availability of semantic processing tools encourage analysts to extend existing link analysis methods towards contextual data processing. To demonstrate presented method, a proof of concept environment IAFEC Ontology Toolkit has been described. It delivers initial financial fraud identification schemes (rules) based on set of problem solving ontologies. The novelty in such approach comes from incorporating heterogeneous types of data, which usually are processed by graph methods. The semantic tool, extend capabilities of graph-based (homogeneous) approach by delivering context-aware indirect association identification, and inference path explanation and inspection capabilities. Presented material describes the method and analytical algorithms, which demonstrate description logic reasoning and graph-based semantic association identification and ranking. Developed method has been implemented, as a Protégé OWL 5.0 environment extension, supplemented with web-services delivering distributed data processing, aggregation (not available in ontology languages). The environment provides declarative processing capabilities enabling analysts to design configurable processing flow, and new financial fraud identification schemes.

1 Introduction and main concept

The IAFEC project, carried out at Military University of Technology for General Inspector of Financial Information (GIIF) aims at developing an analytical method for automatic financial fraud detection. The paper describes the proposed approach while also presenting several similar solutions found in literature.

Increasing computational power combined with new reasoning methods and algorithms provide analytical echanisms applied in many domains. Financial crimes and money laundering are becoming one of key problems for governments and international organisations. The problems mainly concern tax evasion as well as supporting terrorism and organized crime. Technology involvement, the globalization of money transfers, and recently cyber currency make it even harder to identify, seek for, recognise for such actions. Governments for many years have been using analytical tools for monitoring and filtering financial transfers in order to detect and evaluate suspected financial operations. Existing analytical tools rely mostly on big data processing schemes, manually implemented rules and expert's knowledge. The recognition of new fraud schemes, which contain symptoms of financial crime or abuse is a complex task and in major cases need to be performed by financial or tax analysts. Such cases of frauds can be camouflaged by chains of actions associated with organizations or individuals, timespan or loosely coupled facts. To identify such cases there exists, a need for automatic filtering and ranking tools, which can perform preliminary phase of processing and seek for suspected operations from streams of financial data.

The complexity of problems, generates also the need for research and development in the domain of automatic data retrieval, classification and inference. Research projects FF POIROT [1], DOGMA [2] discuss domain models as methods for information retrieval. An extended approach is presented in this paper, which concentrates on application of ontology models, structural analysis and quantitative approach [3] [6] [7] to develop and characterise inference schemes on which a set of tools supporting the analyst has been developed [3] [1] [2] [9].

[a] Mariusz Chmielewski: mchmielewski@wat.edu.pl
[b] Piotr Stąpor: piotr.stapor@wat.edu.pl

2 Problem formulation

The IAFEC project combines several analytical methods for financial fraud detection mainly connected with application of network analysis, reasoning and structural analysis. The novelty of presented approach is to apply domain modelling and problem-solving ontologies in data processing chain, in order to identify hidden associations in knowledge base processed by DL, FOL and semantic graph processing reasoners. Application of semantic models gives also another advantage, context-aware relationships implementation, which extend method's multi perspective capabilities.

3 Analytical methods review

This chapter presents noteworthy approaches that have been found in literature during the review of existing solutions while designing quantitative semantic association ranking method for IAFEC project. The major assumption is to consistently apply structural methods enriching them with semantic analysis thus extending ordinary weighted graph approaches, which enables

3.1 FF Poirot ontology

FF Poirot project was directed against VAT and investments frauds in EU. The inference was done with the help of multilingual ontology integrating knowledge from fields of European law, preventive practices and frauds data. The ontology itself was created with the help of natural language processing mechanisms employed to extract knowledge from unstructured (e.g. legal acts) or poorly structured (e.g. XML documents, database texts) sources. The knowledge acquisition process was automated with a V.U.B. STAR Lab's approach called DOGMA [16]. FF Poirot ontology is logically divided into following layers: hypothesis layer, rules-of-the-law layer and evidence layer. The last one contains decomposition and generalization relations. What's more, evidences can be assessed with a weight that tells to what extent they confirm or reject a given hypothesis. The implemented inference mechanism works in two ways. First one finds a connection from a hypothesis to an evidence. System tries to match the set of laws that must or may be broken if a hypothesis turns out to be true. Than is searches for evidences indicating infringement of laws from the previously obtained collection or lack of it. The second finds a connection from an evidence to a hypothesis. System relies on given evidences to infer which laws have been or have not been broken. On that basis, it tries to determine what abuses have been committed (hypothesis). [12]

3.2 Wigmore's charts

The side result FF Poirot project was also a module intended for automatic e-mail fraud detection. Inference was done with the Wigmore's diagram method (also known as Wigmore's charts). It concentrates on organizing of legal evidences in structures that facilitate the determination of the truth level of a given hypothesis. The unquestionable Wigmore's diagrams' advantage is the possibility automated inference with algorithms taken from the fields of probabilistic theory, fuzzy sets or the of networks' flow theory. [13]

3.3 MEBN

Another interesting idea - Multi-Entity Bayesian Networks (MEBN) - was used in a system deployed in Brazil dedicated to automatic detection of corruption in public procurement processes. The project creators used the descriptive logic enriched with mechanisms that take into account uncertainty. To build the knowledge base a PR-OWL ontology was used, which extends the capabilities of OWL with elements of probability theory. PR-OWL uses the MEBN model which divides the world description (MTheory) into a set of fragments called MFrags, each one containing data about connections that occur between knowledge base individuals and: a) their state (*hasPossibleValues* relation), b) probability distribution (*hasProbDist* relation) and c) other individuals. The implementation was based on UnBBayes-MEBN software, which automatically converts the description contained in the MTheory fragments into the corresponding so-called Situation Specific Bayesian Network (SSBN). Situations for which the obtained results have exceeded certain threshold values are automatically reported, along with the evidence collected by the system, to the relevant decision makers. [14][15]

3.4 StarSoft approach

StarSoft company offers a solution designed for detection of (among others) ATM scams and money laundering practices. The system analyzes historical and current banking transactions using rule inference method performed by SEWSS (STATISTICA Enterprise Wide SPC System). The rules were auto-generated using STATISTICA Data Miner. Detection of relationships between transactions and persons (physical and legal) is achieved using methods from graph and network theory. The structure of the graph is tested for the existence of critical paths and subgraphs that match the fraud patterns. The StarSoft software solution has been designed to learn from mistakes - feedback from an expert automatically improves the classifier's capabilities. Associating similar cases of financial fraud takes place through the use of Kohonen's neural networks, each topological map cell containing suspicious events with similar characteristics. In addition to the mechanism described above, the solution explores, also in a network-based manner, the behavior of the bank account user, capturing suspicious behavior such as change of living place within a short after opening an account. [17]

4 Method outline

To deal with the problem a following solution has been devised. Incoming stream of transactions and alerts (untrustworthy transaction flag) reported from trusted financial institutions are fed into intermediate database, which holds all elementary data about transactions. Fortunately, the gathered data possess a structure in form of schema described in an officially released XSD file. The acquired data should then be transformed into ontology individuals and semantic relations between them so that it may be processed further by HermiT reasoning engine. For that purpose, a GIIF ontology (resembling the mentioned XSD schema, albeit in semantic graph structure) was developed – it acts as a mapping (alignment method) between database and main reasoning ontology. Employing HermiT (or any other semantic reasoner for that matter) with both complex DL constructs and SWRL rules on such a large dataset is costly in terms of time and memory. This has been solved by running a set of aggregation procedures which increases further data processing and inference efficiency. These algorithms have been implemented in Java programming language. The rules they embody, search for certain structure patterns and generate new individuals, each one representing a set of aggregated

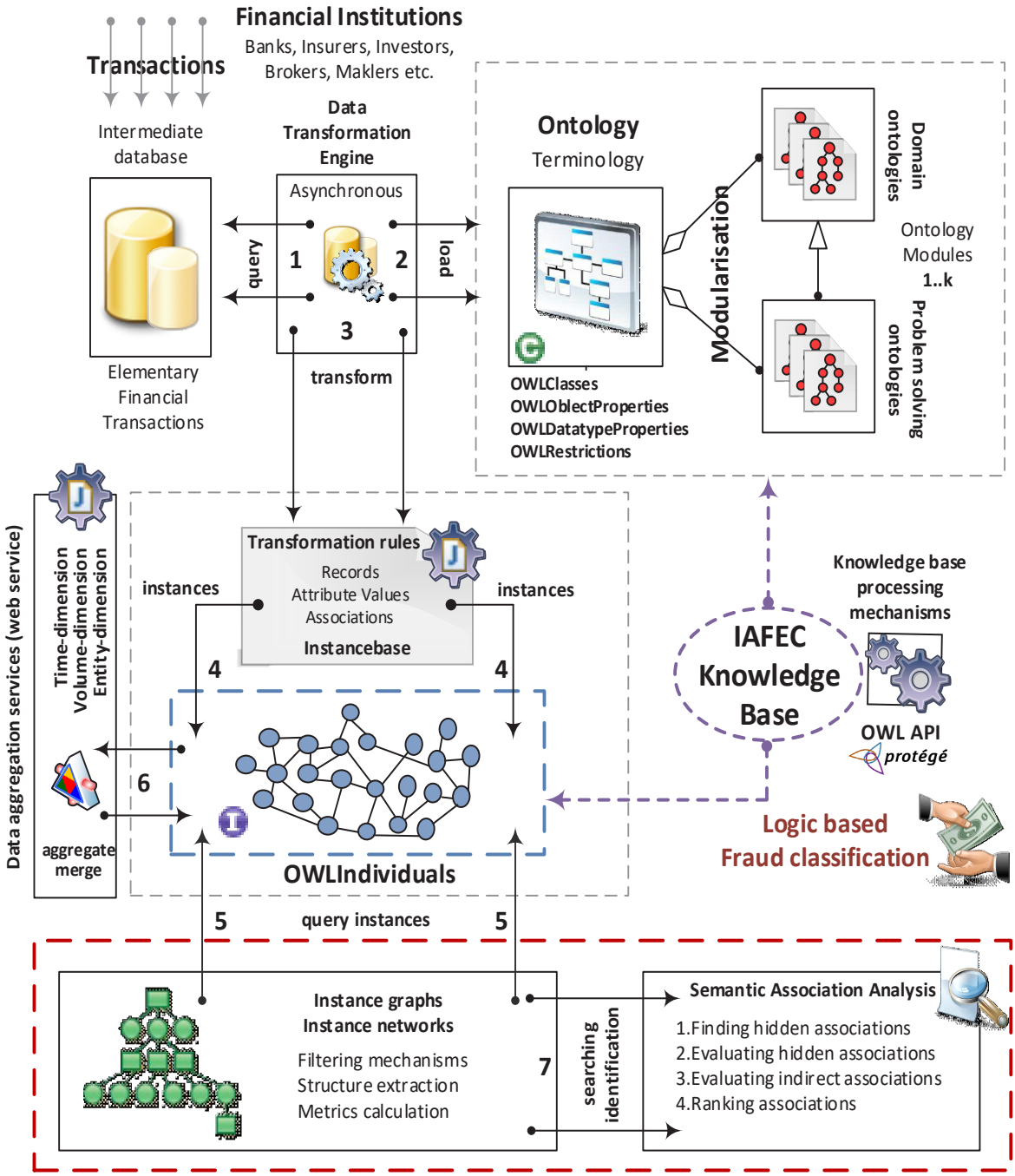

Fig. 1. The method knowledge processing flow – starting from elementary data, semantic data enrichment, aggregation, quantitative evaluation to reasoning. The heterogeneous data and context-aware relationships processing is a result of 1-7 stages of processing power combining logic reasoning and data aggregation.

data and each one equipped with new datatype properties links. These arcs point to values obtained by applying various math functions for individuals' datatype values in the analysed set. E.g. the algorithm sums all transactions done by a given company (represented as datatype relations) in a month and produces as a result an object of OWL class *HighValueTransactionSet* with values such as total and average transaction values.

The aggregated data fuels reasoning engine of main fraud detecting ontology, during which HermiT classifies individuals into concepts by executing its hyper-tableau algorithm and connects (via object-type properties) them according to our set of DL-safe SWRL-defined rules. The end of HermiT's work marks the start of semantic structure analysis. A devised, dedicated tool transforms instances' base into a multigraph and then divides it into a set of weak connected components. of which each one undergoes structural reasoning separately. The tool evaluates various coefficients for vertices, edges and directly not connected pairs of individuals (given, that they belong to the same weak connected component). This approach enables to greatly reduce this stage's running time. Of course, a pair of vertices from different components can in reality be associated, but a probability that such semantic linkage exists is considerably negligible given problem's size and complexity. [7]

Because of resulting multigraphs' magnitude, it might be both difficult and time consuming for user to consume the entire output. It is natural to expect that within such large structure the importance levels of some discoveries prevail over others. The assignment of certain company's instance to a *HighlySuspiciousCompany* concept is undoubtedly more significant than classification of *TransactionSide* as *FinancialInstitution*, but the problem that remains is: how to bring to an attention of a user some meaningful connections realized as chains of instances joined by object-properties links. One of possible ways to deal with it is by the application of *semantic association ranking* method, described in [3]. Please see chapter 5.2 for more details.

5 Semantic descriptions utilised in association detection

5.1 Different perspectives of semantic model structure

The presented method requires to look at semantic model from few perspectives. Let us discuss its structure of ontologies as a first one. The knowledgebase schema consists of layers containing logically and thematically consistent modules. This approach has been chosen for the sake of memory optimisation and clarity.

The GIIF ontology, which act as a mediator between GIIFS XML data format and reasoning module. The latter contains base ontology, which defines primitive concepts used by second layer ontologies: one describing legal persons (institutions, companies, foundations, etc...) and second – describing natural persons in a context of their family relation. On the top lies fraud detecting modules, which analyses transaction chains. During project a few optional modules have been devised: a) value ontology (defines value type hierarchy that includes: financial instruments, material goods, property, services, information, etc...), b) taxonomy of economic sectors to which modelled enterprises may belong, c) extended taxonomy of the money transfer/payment methods (various forms of card payments (e.g. done with debit-, credit-, prepaid- card, online money transfer, material money, loans, withdrawals). OWL language and DL-safe SWRL were used to describe rules for classification of individuals, but only SWRL's Horn clauses could be used to infer about new connections that might materialize between instances.

The second perspective through one may look at a semantic model is a set of concepts' and object-properties' hierarchies.

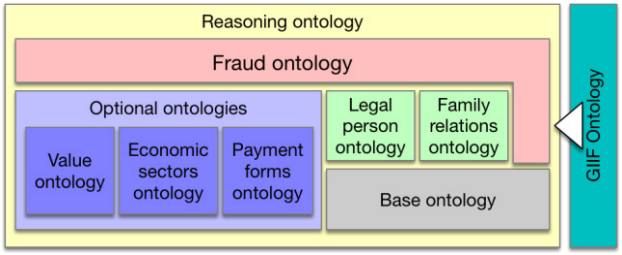

Fig. 2. Diagram shows IAFEC's project ontology modules and layers.

This structure (usually bears slight resemblance to a forest) can be used to analyse validity of subsumption relation, but what is more important, it allows us to automatically assign a numerical value to each class and object-property type in a way that it reflects its informative power and importance. (See [3] for more about structural approach to ontologies.) The more general the concept the less informative power it holds – e.g. a conclusion that an individuum is an instance of TransactionSide concept gives us less knowledge than that it is of a ForeignCompanyWithSubsidiaries type.

In order to express a level of suspicion, some concepts possess three or more grading subclasses – e.g. Company, CompanyWorthConsidering, Suspicious-Company, HighlySuspiciousCompany and SurelyFraudulentCompany. The lower the position in hierarchy the more important and noteworthy is the discovery. After DL and SWRL reasoning completes, the third ("structural") perspective can be applied. It views the resulting instance base as a multigraph.

This enables us to employ various methods form graphs and nets theory. E.g. link relevance (also known as Jaccard's index/coefficient) can infer a connection between two entities by counting their common neighbours and dividing the result by number of all their neighbours.

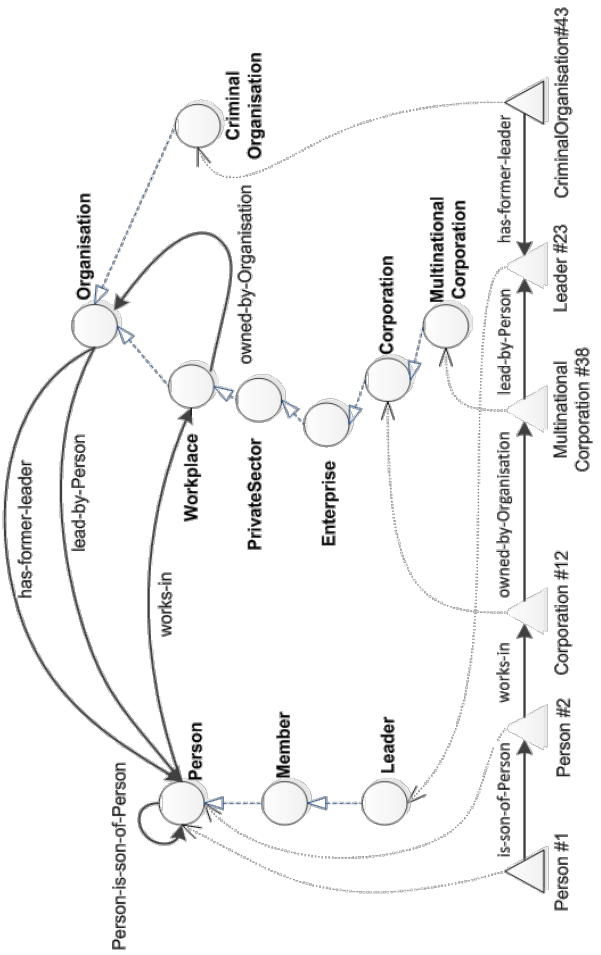

Fig. 3. Representation of semantic association – upper part describes terminology and lower data instances. Quantitative approach assigns instance measures based on instance type measures.

5.2 Semantic associations

In order to assess importance factor of long sequences, a notion of semantic association [3] could be employed. The semantic association pattern is a chain of intertwined concepts and object-properties. The reasoning engine's goal is to find (in instance base) sequences of individuals and links between them such that they belong respectively to (sub)concepts and (sub)properties defined in the given pattern. The importance of such semantic association (chain) can then be expressed with the help of values obtained from both hierarchical (and in some cases structural) perspective in a manner that promotes nodes and links from lower parts of hierarchy. [3]

To better illustrate the idea figure 3 showing association has been included. The tables represent values obtained by structural analysis. E.g. Concept Capacity Taxonomy Measure (CCTM) for a given concept c is evaluated as the length of the longest c to owl:Thing path in concepts hierarchy divided by the length of the longest path in entire hierarchy containing c.[3] for definitions of measures.

Concept instance	CCTM	ACCTM	SCCM (CTG)	ACSD	SCBM	CLNCC	SCCM (CTG) ACSD	SCCM (CTG) SCBM	Concept Score
Organisation #1	0,2857	0,2857	0,5000	0,5921	0,9175	0,0476	0,6092	0,6305	0,42237
Enterprise #1	0,7143	0,7143	0,0714	0,0263	0,0000	0,0000	0,0345	0,0559	0,37448
Stakeholder #20	1,0000	0,4286	0,2000	0,0000	0,0000	0,0000	0,0000	0,0000	0,43429
Person #322	0,3333	0,2857	0,5000	1,0000	1,0000	0,0778	0,9885	1,0000	0,56225
CyberTerrorist #23	1,0000	0,7143	0,0714	0,0000	0,0000	0,0000	0,0000	0,0000	0,46429
CrimeEvent #53	0,7500	0,4286	0,2000	0,0000	0,0000	0,0000	0,0115	0,0135	0,34862
Organisation #3	0,2857	0,2857	0,5000	0,5921	0,9175	0,0476	0,6092	0,6305	0,42237
Organisation #52	0,2857	0,2857	0,5000	0,5921	0,9175	0,0476	0,6092	0,6305	0,42237
Person #22	0,3333	0,2857	0,5000	1,0000	1,0000	0,0778	0,9885	1,0000	0,56225
Measure weights	*0,35*	*0,15*	*0,1*	*0,15*	*0,05*	*0,05*	*0,1*	*0,05*	**4,01327**

Fig. 4. Multi-criteria semantic association ranking, evaluation of instance data based on their terminology importance measures: CCTM - Concept Capacity Taxonomy Measure, ACCTM - Absolute CCTM, SSCM - Structural Concept Closeness Measure [3]

First table covers instances, while second presents figures for object-properties. The lowest row in both cases defines the weight the given measure plays in assessment of an element. The last column holds a weighted sum of these measures for each instance or link. The lowest, leftmost cell shows a sum of all other values in its column, which may be interpreted as final score for the association.

In the next step a two-argument function (e.g. average, sum) could be utilized, that would bind both results (for instances and links) into a final semantic association's importance/informative estimation.

Structural relationship instance	ASRD	ASRD in	SRSCTM	ASRCTM	SSRCM	SMSRCM	Relation Score
(Org) associated-with (Org)	0,0196	0,5000	0,3333	0,1667	1,0000	1,0000	0,44461
has-elected-leader	0,0000	0,5000	1,0000	0,5000	0,1667	0,3333	0,48334
is-father-of	0,0000	0,5000	1,0000	0,8333	0,0667	0,2000	0,52167
is-friend-of	0,0196	0,5000	0,6667	0,3333	0,3333	0,5000	0,41127
caused-by	0,0196	0,5000	0,6667	0,6667	0,1000	0,2500	0,41796
harmed-in	0,0000	0,5000	1,0000	0,6667	0,1000	0,2500	0,49834
(Org) associated-with (Org)	0,0196	0,5000	0,3333	0,1667	1,0000	1,0000	0,44461
is-founder-of	0,0000	0,5000	1,0000	0,1667	1,0000	1,0000	0,60834
Measure weights	*0,15*	*0,15*	*0,25*	*0,2*	*0,15*	*0,10*	**3,83012**

Fig. 5. Multi-criteria semantic association ranking, evaluation of structural relations importance measures: ASRD (in) - Absolute Structural Relationship (in) Degree, SRSCTM - Structural Relationship Semantic Capacity Taxonomy Measure. See [3] for details.

5.3 Transaction data context

The proposed semantic model augments transactions layer with additional data. Our approach enables to look at financial flows in context (from perspective) of organizations' dependencies, workplace relation-ships and even family-relations. In this way, system takes the advantage of the fact that its user – GIIF – is a government establishment and thus possesses ability to retrieve data from other polish institutions such as CEPiK, CEIDG, KRS to increase the probability of fraud detection. Among features that trigger suspicion one can find: a) frequent exchange of financial instruments (hence value ontology module); b) cycles of transactions between subsidiaries; c) mutual loans; d) company keeps making transactions that surpass its value for a longer period of time.

5.4 Aggregation scheme example

Fig. 6. presents a fragment of sample instance base graph containing three transactions.

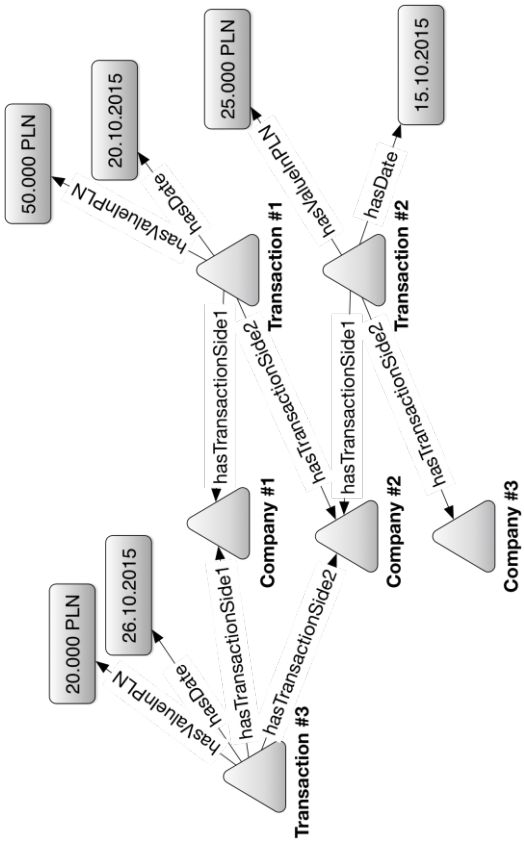

Fig. 6. An example of transactions' instances graph structure before aggregation.

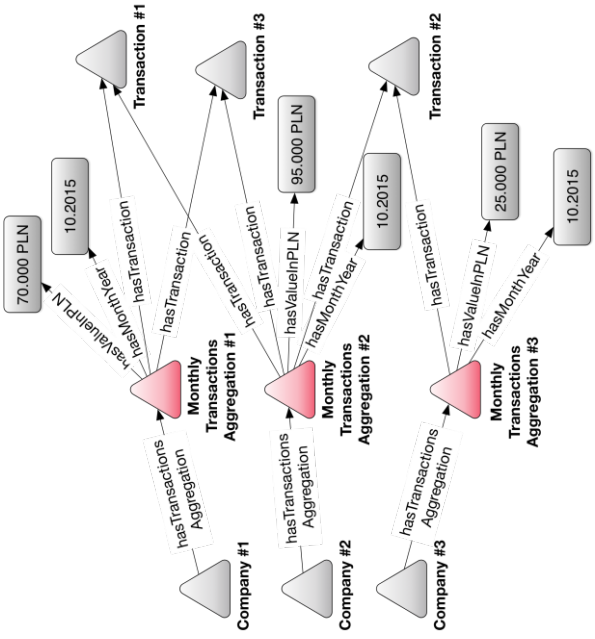

Fig. 7. Aggregation procedure produces monthly values' flow summary for each company.

It is only after the aggregation procedure (Fig. 6. - Fig. 7.) produces monthly summary nodes for each company, new information such as Company #2 having the biggest money flow in October 2015 could be extracted.

However, description logic cannot be used to perform arithmetic operations on datatype-properties, neither DL safe rules allows that. That is the reason why aggregation procedures had to be implemented as separate Java components.

The figure below shows: a) new generated instances that represent transactions' summaries; b) automatically added connections (using object-properties) that bind these new nodes with already existing individuals; c) auto-inserted datatype-properties' arcs pointing to calculated numerical values.

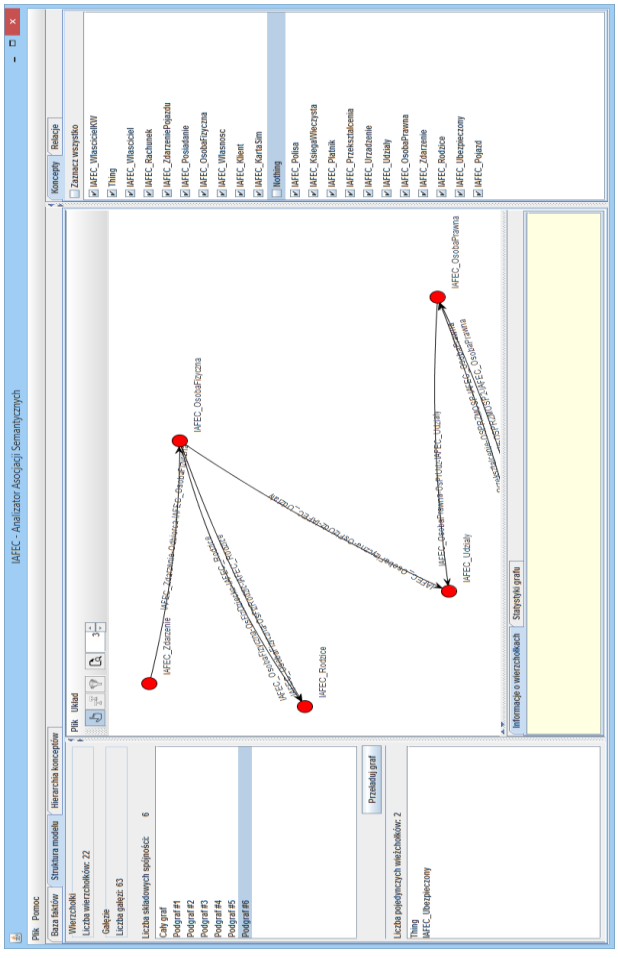

Fig. 8. IAFEC Semantic Association Analyzer main window showing fragment of structural multigraph.

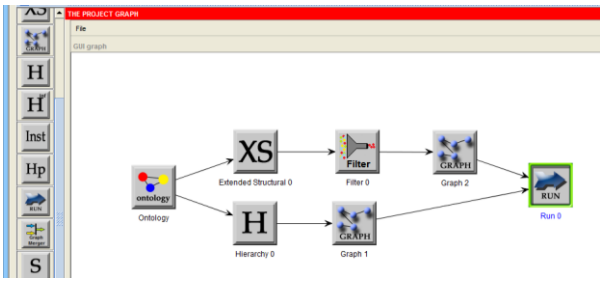

Fig. 9. IAFEC Semantic Association Analyzer calculation process designer delivering features for semantic model evaluation based on analyst process definitions

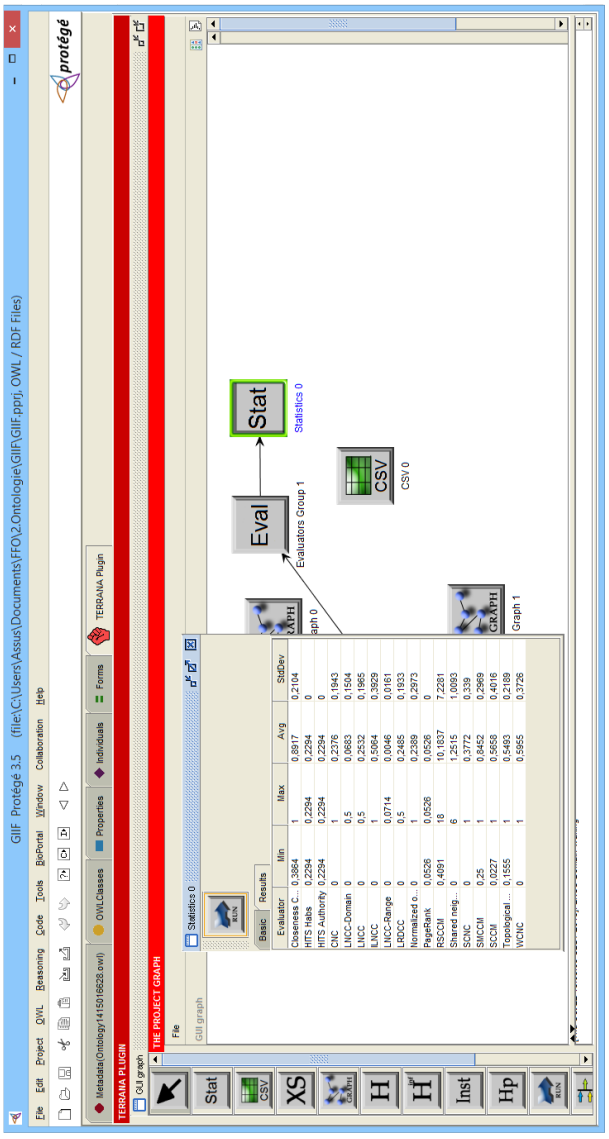

Fig. 10. IAFEC Semantic Association Analyzer integrated with TERRANA plugin demonstrating structural characteristics evaluation of semantic models – conducted for GIIF ontology family

6 Evaluation environment

The method required to use a number of available tools. The ontologies were developed and tested using Protégé 5 environment with HermiT reasoning engine. Our proof of concept included implementation of dedicated, server side, data aggregation module. This was achieved with the help of owl-api library. Aggregation methods could be invoked on-demand via supplied webservice. Ultimately, a dedicated tool was implemented, that enabled to divide resulting multigraph structure into weak connected components and visualize it with the help of jung graphing library while offering application of various filters. The structural analysis was performed with the help of reused modules of a next version of Terrana environment [9]. The tool allowed us to fuse values acquired from hierarchical perspective with structural data. More information about the dedicated application - IAFEC Semantic Association Analyzer tool can be found in [7].

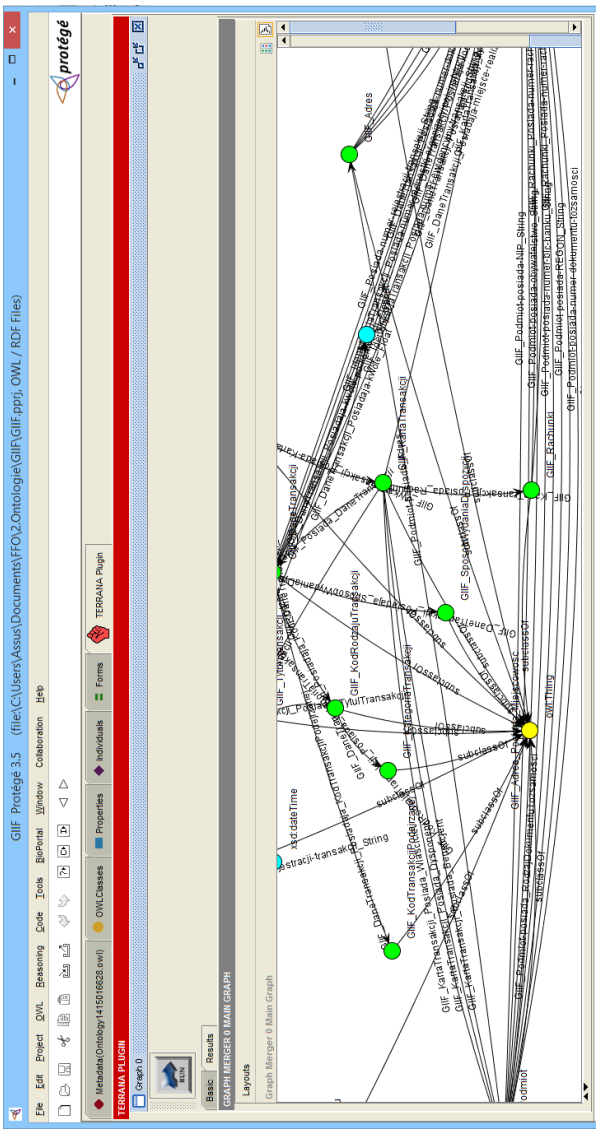

Fig. 11. Semantic instance base graph (financial transactions data) analysed in TERRANA plugin prepared for filtering and quantitative structural analysis - conducted for GIIF and IAFEC ontology families

7 Summary

Presented method and tool delivers defined quantitative approach for processing semantically relevant financial data, which can be used to identify and recognise, complex financial abuses. Encasing the method in form of analytical software such as IAFEC Semantic Association Analyzer tool, delivers declarative analytical flow capabilities. Tool can be integrated with preliminary filtering services of large scale data. In the second stage of processing, the described tools implement human-in-the-loop algorithms, which deliver configurable and adjustable analytical schemes. Such construction is intentional as the environment should deliver capabilities for seamless integration within larger heterogeneous SOA environment. The framework architecture of IAFEC environment provides, a process based analytical workflow, which integrates various methods. Presented method delivers complete set of analytical mechanisms for data transformation, aggregation, filtering and evaluation. Network-based analysis, graph querying are effective tools for specific data processing, thus introduction of semantic descriptions and reasoning extend these capabilities. Scenario-based analysis as well as in-stream data selection and filtering are valuable tools for analysts which after testing the usefulness of described scheme can deploy the rules for large scale processing. During software functional tests, a set of financial fraud datasets have been analysed, providing expected identification results. However, the analyst needs to consider the limitations of ontology processing model relying on in-memory representation. Most recent development of IAFEC semantic tools move towards utilizing Neo4J graph database and triple stores in order to eliminate memory limitations. To support effective ontology model management, a modularization methodology has been implemented, which simplifies memory management and supports instance data loading during the analysis. This supports the knowledge engineer while extending terminology and delivers larger instance base processing capabilities.

Acknowledgements: This work was supported by National Centre for Research and Development project DOBR/0073/R/ID1/2012/03: "Advanced ICT techniques supporting data analysis processes in the domain of financial frauds"

References

1. Basili R. et al.: Ontology-driven Informa-tion Retrieval in FF-Poirot. (2003)
2. Jarrar M., Meersman R.: Ontology Engineering - The DOGMA Approach. In: Advances in Web Semantics I. LNCS, vol. 4891, Springer (2008)
3. Chmielewski M.: Ontology-based association assessment method using graph and logic reasoning

techniques. Ph.D. thesis, Military University of Technology, Warsaw (2011)
4. Ministry of Finance Republic of Poland, http://www.mf.gov.pl/ministerstwo-finansow/dzialalnosc/giif/system
5. Dentler K. et al.: Comparison of Reasoners for large Ontologies in the OWL 2 EL Profile. In: Semantic Web, vol. 2 (2011)
6. Chmielewski M., Stąpor P.: Medical Data Unification Using Ontology-Based Semantic Model Structural Analysis, Proceedings of 36th International Conference on Information Systems Architecture and Technology, (2016)
7. Chmielewski M., Stąpor P. Money Laundering Analytics Based on Contextual Analysis. Application of Problem Solving Ontologies in Financial Fraud Identification and Recognition, Information Systems Architecture and Technology – ISAT 2016 – Part I. Advances in Intelligent Systems and Computing, vol 521. Springer, Cham, (2017)
8. Barthelemy M. et al.: Knowledge Representation Issues in Semantic Graphs for Relationship Detection. In: AAAI Spring Symposium: AI Technologies for Homeland Security. (2005)
9. Chmielewski M., Stąpor P.: Protégé based environment for DL knowledge base structural analysis. In: Computational Collective Intelligence. Technologies and Applications (2011).
10. Chmielewski M., Paciorkowska M., Kiedrowicz M., A semantic similarity evaluation method and a tool utilised in security applications based on ontology structure and lexicon analysis, 21st International Conference on Circuits, Systems, Communications and Computers (CSCC 2017), Greece, July 14-17, 2017
11. Kiedrowicz M., et al., Business process data flow between automated and human tasks, 3rd International Conference on Social Science (ICSS 2016) December 9–11 2016, pp. 471-477, (2016).
12. Leary R.M., et al., Towards a financial fraud ontology; a legal modelling approach. In Proceedings of the ICAIL 2003 Workshop on Legal Ontologies & Web based legal information management, 2003.
13. Chalamish M, et al. Intelligent evaluation of evidence using Wigmore diagrams. In Proceedings of the 13th International Conference on Artificial Intelligence and Law, ICAIL '11, pages 61–65, New York, NY, USA, 2011. ACM.
14. Rommel N Carvalho et al. Probabilistic ontology and knowledge fusion for procurement fraud detection in Brazil. URSW, 527:3–14, 2009.
15. Kathryn B Laskey et al. Pr-owl 2 case study: A maritime domain probabilistic ontology. In STIDS, pages 76–83, (2011).
16. Peter Spyns et al. Evaluating dogma-lexons generated automatically from a text corpus. STAR, 2004(13):13, 2004.
17. Ton Kuijlen and Grzegorz Migut. Wykrywanie nadużyć i prania brudnych pieniędzy. Stat Soft, pages 71–80, 2004.