RDF Data Analysis with Activation Patterns

Peter Teufl
(IAIK, Graz University of Technology, Graz, Austria
peter.teufl@iaik.tugraz.at)

Günther Lackner
(studio78.at, Graz, Austria
guenther.lackner@studio78.at)

Abstract: RDF data can be analyzed with various query languages such as SPARQL or SeRQL. Due to their nature these query languages do not support fuzzy queries. In this paper we present a new method that transforms the information presented by subject-relation-object relations within RDF data into *Activation Patterns*. These patterns represent a common model that is the basis for a number of sophisticated analysis methods such as semantic relation analysis, semantic search queries, unsupervised clustering, supervised learning or anomaly detection. In this paper, we explain the *Activation Patterns* concept and apply it to an RDF representation of the well known *CIA World Factbook*.

Key Words: machine learning, knowledge mining, semantic similarity, activation patterns, RDF, fuzzy queries

Categories: M.7

1 Introduction and Related Work

In the semantic web knowledge is presented by the Resource Description Format (RDF), which stores subject-predicate-object triplets (e.g. in XML format). An example for such a triplet would be the fact that Austria (subject) has the Euro (object) as currency (predicate). An RDF data source[1] can therefore describe arbitrary aspects of arbitrary resources and is an example for a semantic network. Since subjects, predicates and objects are identified via unique URIs[2], various RDF sources can easily be merged. In order to extract information from RDF data, various query languages such as SPARQL [W3C(2008)] or SeRQL [Broekstra and Kampman(2004)] have been developed. In SeRQL the query *"SELECT countries FROM (countries) border (Germany)"* selects all *countries* (subject) that *border* (predicate) *Germany* (object)[3]. Such query languages allow the retrieval of arbitrary information stored within the RDF data source. However, they do not allow us to find answers for questions like

[1] e.g simple XML files, databases etc.

[2] Objects can either be values (e.g. Strings or real values) or refer to other subjects identified by URIs.

[3] The employed XML namespaces are not shown in the query.

"How is the literacy rate typically related to the unemployment rate? Retrieve all countries according to their similarity to Austria; Find the typical features for countries that export bananas and retrieve all countries that have similar features but do not export bananas themselves; Group countries according to feature values related to gross domestic product and imported commodities." All of these queries are fuzzy in their nature and query languages such as SPARQL or SeRQL cannot directly be used to find the answers.

Therefore, we present the concept of *Activation Patterns*. The basic idea is to represent knowledge and its relations within a semantic network (RDF data sources are semantic networks). *Activation Patterns* are then generated by activating nodes (subjects) and spreading this activation over the network (through predicates). The *Activation Pattern* is a vector representation of the node activation values within the semantic network. These patterns allow the application of a wide range of fuzzy analysis methods to arbitrary RDF data sources.

The paper gives an introduction on the concept of *Activation Patterns* and shows the possibilities by applying the technique to an RDF representation of the *CIA World Factbook*[4].

As our work is part of the broad field of semantic searching, a detailed description of related work would go far beyond the scope and space limitation of this article. The general idea of using patterns to search semantic networks is fairly old and has among others been formulated by [Minker(1977)] 30 years ago. In the following years various approaches have been developed and described i.e. by [Crestani(1997)] and [Califf and Mooney(1998)]. Statistical and graph based methods have mainly been in the focus of past research work. The current movement back towards AI based techniques, where our work is aligned with, promises further improvements in performance and reliability as these techniques significantly evolved in the recent years [Halpin(2004)]. Interested readers are requested to refer to general literacy and the following references: [Lamberti et al.(2009)], [Kim et al.(2008)], and [Ding et al.(2005)].

2 Activation Patterns

The following explanation is based on features and relations taken from the *CIA World Factbook*, which will be analyzed in the last section. Assuming we want to learn more about countries (subjects), we need to generate patterns for each country instance. The transformation of these instances into *Activation Patterns* is based on five process layers depicted in Figure 1. After extracting and pre-processing the country features, we apply the four layers L1-L4

[4] https://www.cia.gov/library/publications/the-world-factbook/

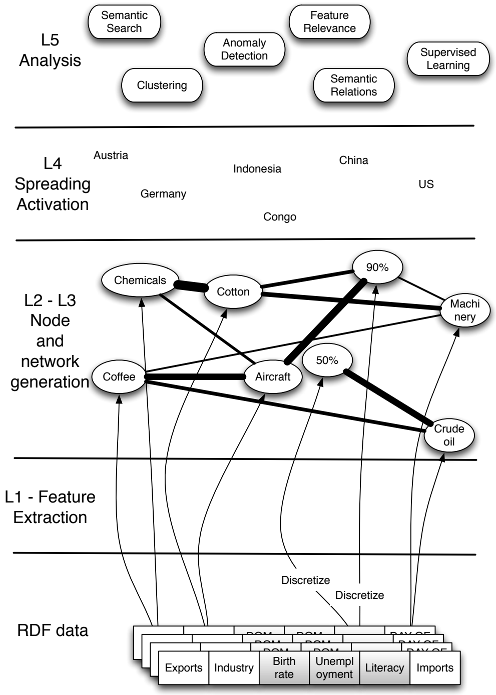

Figure 1: Layers for Activation Pattern generation

to the raw feature data in order to generate the *Activation Patterns*[5]. The techniques within these layers are based on various concepts related to machine learning and artificial intelligence: semantic networks for modeling relations within data [Quillian(1968)], [Fellbaum(1998)], [Tsatsaronis et al.(2007)], spreading activation algorithms (SA) [Crestani(1997)] for extracting knowledge from semantic networks, and supervised/unsupervised learning algorithms to analyze data extracted from the semantic network [Martinetz and Schulten(1991)], [Qin and Suganthan(2004)]. The basic idea is to create a semantic network that stores the feature values and the relations between these feature values (L2-L3). The *Activation Patterns* are then generated by applying spreading activation algorithms (L4) to the semantic network. Such a pattern is the vector representation of all the activation values of the semantic network nodes. The *Activation*

[5] Since RDF data is already a semantic network, we would not need L1-L3 (or only parts of them) for the pattern generation. However, since our framework is not completely adapted to RDF data yet, we still need these layers for the generation process.

Pattern concept allows us to apply various analysis techniques in L5.

Semantic relations: The analysis of the nodes and the links within the semantic network allows us to gain knowledge about the relations within the given data. For example, by specifying a certain *literacy* value we are able to find how strong the other features such as *unemployment rate* or *gross domestic product* are related.

Semantic search: Such search queries utilize the links (relations) within the semantic network to find concepts related to the search query. For example, by specifying an *export commodity* (e.g. *crude oil*) we are able to retrieve similar countries even if they do not export this *commodity* but are otherwise related.

Unsupervised clustering: By grouping (clustering) semantically related patterns, countries with similar features are assigned to the same category. An example would be the creation of categories that group countries according to similar *export partners*. The number of clusters (or model complexity) influences the grade of detail covered by each category.

Supervised learning: If class labels are available, supervised learning algorithms can directly be applied for the training and classification of *Activation Patterns*.

Feature relevance: The relevance of a node representing a feature value within the network can be determined by the number of connections from this node. Nodes with a high number of connections carry less information than those with fewer connections.

Anomaly detection: By analyzing the activation energies of given instances, anomalies can be detected. In the area of network security this plays an important role for the detection of unknown attacks.

We have also applied the *Activation Patterns* concept to other domains such as event correlation for IDS systems, [Teufl et al.(2010)], e-Participation [Teufl et al.(2009)] and malware analysis.

3 CIA World Factbook RDF analysis

In order to show the benefits of the *Activation Patterns* concept we have analyzed an RDF representation[6] of the *CIA World Factbook*[7]. Since the features used to describe the countries are well known, this RDF dataset is a perfect choice for demonstrating and evaluating the *Activation Pattern* concept. In this section we extract features for each country by utilizing the SESAME framework[8] and the SeRQL query language. We then generate *Activation Patterns* for all countries

[6] http://simile.mit.edu/wiki/Dataset:_CIA_Factbook

[7] https://www.cia.gov/library/publications/the-world-factbook/

[8] http://www.openrdf.org/

and use the patterns for the application of three different analysis methods: semantic relations, semantic search queries and unsupervised clustering[9].

3.1 Relations between Objects

Table 1 shows examples for extracting information about semantic relations. For each feature we take all nodes and norm their activation values with the maximum value. Therefore the strongest value is always 1.0.

Relation 1 - *"How do the typical values for unemployment rate, literacy and gross domestic product sectors compare between Africa and Europe?"* By activating the node for *Africa*, we are able to extract the feature values that are typical for countries on this continent. The results for *Africa* are shown in column 1 and 2. Countries within *Africa* typically have a high *unemployment rate*, a rather low *literacy* rate and a large part of the work force is within the *agriculture sector.* The most common exports are *coffee* and *cotton.* Columns 3 and 4 show the results for European countries, which are – as expected – rather different.

Relation 2 - *"How do the same values compare for countries that export crude oil vs. countries that export machinery, equipment and chemicals?"* Here we can see that countries of the second category are typically better developed than countries of the first category (e.g. unemployment rate, literacy, services).

3.2 Semantic Aware Search Queries

By comparing *Activation Patterns* with a distance-measure (e.g. cosine similarity) we are able to find patterns that activate similar regions on the semantic network and are therefore related. This enables us to select an existing pattern (e.g. for *Austria*) and search for related countries. Furthermore, we can execute semantic search queries that only select some of the features. In this case we create an *Activation Pattern* for the given features and values and compare this generated pattern with the existing patterns.

In **Query 1** we execute a query that corresponds to *"List all countries according to their similarity with Austria"*: Therefore the *Activation Pattern* for Austria is taken and compared to the patterns of all other countries by utilizing the cosine similarity as distance measure: best machting: Germany, Sweden, Switzerland, Netherlands, worst matching: Sudan, Gaza Strip, West Bank.

In **Query 2** (see Table 2) we want to *"Find the typical features for countries that export crude oil and retrieve all countries that have similar features but do not export crude oil themselves"*: Therefore, we activate the *crude oil* node, generate the *Activation Pattern* and search for similar country patterns. The

[9] Currently, we use Matlab for all processing steps. However, a more sophisticated Java library is under development.

Relation 1	*mapReference: Africa*		*mapReference: Europe*	
unemploym. (%)	24.45 (1.0)	52.87 (0.4)	04.02 (1.0)	12.93 (0.7)
literacyTotal	41.75 (1.0)	59.85 (0.9)	95.92 (1.0)	80.06 (1.0)
grossAgriculture	40.41 (1.0)	14.80 (0.8)	03.68 (1.0)	14.80 (0.2)
grossServices	37.45 (1.0)	48.77 (0.8)	70.37 (1.0)	60.15 (0.8)
grossIndustry	21.24 (1.0)	30.39 (0.7)	30.39 (1.0)	21.24 (0.4)
exports	coffee (1.0)	cotton (0.8)	chemicals (1.0)	machinery and equipment (0.7)
Relation 2	*exports: crude oil*		*exports: machinery, equipment* *exports: chemicals*	
unemploym. (%)	24.45 (1.0)	04.02 (0.2)	04.02 (1.0)	12.93 (0.9)
literacyTotal	80.06 (1.0)	95.92 (1.0)	95.92 (1.0)	80.06 (0.1)
grossAgriculture	03.68 (1.0)	14.80 (0.6)	03.68 (1.0)	14.80 (0.3)
grossServices	48.77 (1.0)	37.45 (0.7)	70.37 (1.0)	60.15 (0.8)
grossIndustry	41.40 (1.0)	53.26 (1.0)	30.39 (1.0)	41.40 (0.3)
exports	crude oil (1.0)	coffee (0.1)	machinery and equipment (1.0)	chemicals (1.0)

Table 1: Relations for given features and values. Only a fraction of available features is shown in the table. For each feature the two strongest values are taken. Due to the employed max-norm the strongest value is always equal to 1.0.

results for the 11 best matching countries are not shown, since they are *crude oil* exporters. They could have been retrieved with simple keyword matching (= *crude oil*). More interesting are the results that contain countries that do not export *crude oil* but are still related to the countries which do (results 12 to 16). Although they do not share the value *crude oil*[10] they have similar industries, export goods and other features. Result 202 (at the end of the list) shows a country that is not typical at all for a *crude oil* exporter – *Germany.*

3.3 Unsupervised Clustering

By applying unsupervised clustering algorithms to the *Activation Patterns* of the country instances, we are able to find groups of similar countries. Depending on the focus of the unsupervised analysis we can filter the *Activation Patterns* according to certain features. In the example given in Table 3 only the features

[10] Kuwait lists *oil and refined products* as export commodity. This commodity is not equal to *crude oil*, since they are represented with different nodes within the network. Still, Kuwait is retrieved due to other semantic similarities.

Query 2	**Query for exports:crude oil**	
Result	**Country**	**exports**
12	Equatorial Guinea	timber, cocoa, petroleum
13	Congo	lumber, cocoa, petroleum
14	Kuwait	fertilizers, oil and refined products
15	Cameroon	lumber, cotton, petroleum products
16	Qatar	petroleum products, fertilizers, steel
202	Germany	chemicals, textiles, foodstuffs

Table 2: Semantic search queries

for the distribution of the gross domestic product are taken (percentage: industry, agriculture, services). For clustering we apply the Robust Growing Neural Gas (RGNG) algorithm [Qin and Suganthan(2004)] to the *Activation Patterns*. By utilizing a simple model complexity we get the three clusters shown in the table. Cluster 1 represents countries with a very small agricultural sector (typically rich countries). In contrast Cluster 3 represents those countries with a rather large agricultural part and small services part (typically poor countries). Cluster 2 is somewhere in the middle between Cluster 1 and 3. The table also shows the typical export commodities for the countries within the clusters, which correspond to the gross domestic product sectors.

4 Conclusions and Future Work

In this paper we demonstrate the benefits of utilizing *Activation Patterns* for the analysis of RDF data. Currently, the RDF data needs to be transformed into another semantic network in order to be compatible with our framework. However, for the future we want to avoid this step, since the RDF data already corresponds to a semantic network and therefore could be directly used for the *Activation Patterns* process. Furthermore, we are currently in the process of creating a *Google Web Toolkit*[11] based visualization interface that allows the application of the discussed analysis methods and the visualization of the results.

[11] http://code.google.com/webtoolkit/

Cluster 1	**Feature values**
exports	machinery and equipment, chemicals manufactured goods, metals food products
mapReference	Europe (1.0) , North America (0.3), Oceania (0.1)
grossAgriculture	3.68 (1.0)
grossServices	70.37 (1.0), 60.15 (0.5)
grossIndustry	30.39 (1.0)

Cluster 2	**Feature values**
exports	sugar, coffee, textiles, electricity, chemicals shrimp, lobster, gold, timber
mapReference	Central America and the Caribbean (1.0) Middle East (0.3), South America (0.1)
grossAgriculture	14.80 (1.0)
grossServices	60.15 (1.0), 48.77 (0.7)
grossIndustry	30.39 (1.0)

Cluster 3	**Feature values**
exports	cotton, coffee, cocoa, timber, diamonds fish, aluminum, gold, livestock
mapReference	Africa (1.0), Southeast Asia (0.3) Asia (0.1)
grossAgriculture	40.41 (1.0)
grossServices	37.45 (1.0)
grossIndustry	21.24 (1.0), 30.39 (0.1)

Table 3: Relations for given features and values

References

[Broekstra and Kampman(2004)] Broekstra, J., Kampman, A.: "Serql: An rdf query and transformation language"; (2004).

[Califf and Mooney(1998)] Califf, M. E., Mooney, R. J.: "Relational learning of pattern-match rules for information extraction"; 328–334; 1998.

[Crestani(1997)] Crestani, F.: "Application of spreading activation techniques in information retrieval"; (1997).

[Ding et al.(2005)] Ding, L., Finin, T., Joshi, A., Peng, Y., Pan, R., Reddivari, P.: "Search on the semantic web"; Computer; 38 (2005), 62–69.

[Fellbaum(1998)] Fellbaum, C.: "Wordnet: An electronic lexical database (language, speech, and communication)"; Hardcover (1998).

[Halpin(2004)] Halpin, H.: "The semantic web: The origins of artificial intelligence redux"; (2004).

[Kim et al.(2008)] Kim, J.-M., Kwon, S.-H., Park, Y.-T.: "Enhanced search method for ontology classification"; IEEE International Workshop on Semantic Computing and Applications; IEEE Computer Society, 2008.

[Lamberti et al.(2009)] Lamberti, F., Sanna, A., Demarti, C.: “A relation-based page rank algorithm for semantic web search engines”; IEEE Transactions on Knowledge and Data Engineering; 21 (2009), 1, 123–136.

[Martinetz and Schulten(1991)] Martinetz, T., Schulten, K.: “A ”neural gas” network learns topologies”; T. Kohonen, K. Mäkisara, O. Simula, J. Kangas, eds., Artificial Neural Networks; 397–402; Elsevier, Amsterdam, 1991.

[Minker(1977)] Minker, J.: “Control structure of a pattern-directed search system”; SIGART Bull.; (1977), 63, 7–14.

[Qin and Suganthan(2004)] Qin, A. K., Suganthan, P. N.: “Robust growing neural gas algorithm with application in cluster analysis”; Neural Netw.; 17 (2004), 8-9, 1135–1148.

[Quillian(1968)] Quillian, M. R.: “Semantic memory”; (1968).

[Teufl et al.(2010)] Teufl, P., Payer, U., Fellner, R.: “Event correlation on the basis of activation patterns”; (2010), 0 – 0.

[Teufl et al.(2009)] Teufl, P., Payer, U., Parycek, P.: “Automated analysis of e-participation data by utilizing associative networks, spreading activation and unsupervised learning”; (2009), 139–150.

[Tsatsaronis et al.(2007)] Tsatsaronis, G., Vazirgiannis, M., Androutsopoulos, I.: “Word sense disambiguation with spreading activation networks generated from thesauri”; (2007).

[W3C(2008)] W3C: “SPARQL query language for RDF”; (2008).