Journal of Universal Computer Science, vol. 17, no. 7 (2011), 983-1004
submitted: 30/10/10, accepted: 15/3/11, appeared: 1/4/11 © J.UCS

Knowledge Extraction from RDF Data 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. However, due to their nature these query languages do not support fuzzy queries that would allow us to extract a broad range of additional information. In this article 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 article, 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

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 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 SPARQL the following query selects all *countries* (subject) that have *iron* (object) as *export-commoditiy* (predicate)[3].

[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.

```
PREFIX    foaf: <http://xmlns.com/foaf/0.1/
SELECT    ?country
FROM      <factbook.rdf>
WHERE     { ?country foaf:export-commodity "iron" }
```

RDF query languages allow the retrieval of arbitrary information stored within the RDF triplets. However, they do not allow us to find answers for fuzzy questions like:

– *How do the typical values for unemployment rate, literacy, gross domestic product sectors and exports compare between Africa and Europe?*

– *How do the same values compare for countries that export crude oil vs. countries that export machinery, equipment and chemicals?*

– *List all countries according to their semantic similarity with Austria.*

– *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.*

– *Cluster countries according to feature values related to gross domestic product and exported commodities.*

– *What are the most relevant features used for the description of the countries?*

All of these queries are fuzzy in their nature and typical RDF query languages cannot 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. *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 and enables us to apply a wide range of fuzzy analysis methods to arbitrary RDF data sources.

The article first gives a detailed description of the *Activation Patterns* concept and then shows the benefits by applying the technique to an RDF representation of the *CIA World Factbook*[4].

2 *Activation Patterns* and RDF data analysis at a glance

Activation Patterns are generated by a transformation process that takes an arbitrary combination of *symbolic* (e.g. language) and *distance-based* features (e.g. unemployment rate) as input. The RDF-relations contained in the data

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

set are modeled as nodes and links between these nodes in a semantic network. Knowledge about the stored relations is extracted by activating nodes within the network and spreading the activation to neighboring nodes. The activated network is then represented as a vector – the *Activation Pattern*. These patterns are the basis for a wide range of analysis techniques. We have already successfully deployed the technique within various other domains:

- **Event correlation** plays an important role in network security related areas. Intrusion Detection Systems (IDS) typically collect events from arbitrary sensors. In order to analyze or detect attacks the relations between these events need to be understood. In [Teufl et al.(2010b)] we have applied the *Activation Patterns* concept for the analysis of IDS data.
- **e-Participation** is an area where a large number of text documents need to be analyzed. The *Activation Patterns* concept can also be used to represent documents and the semantic relations between terms [Teufl et al.(2009)]. Furthermore, the evolvement of semantic relations between terms over time can be analyzed.
- **Malware analysis** involves the analysis of machine code (e.g. assembler code or even high level languages such as JavaScript). The *Activation Pattern* concept can be utilized to get a better understanding of the relations between the instructions within the analyzed malware. [Teufl et al.(2010a)]
- **Location tracking** of users in WiFi networks is possible if a behavioral fingerprint of a user can be created. We have employed *Activation Patterns* to discover the most privacy leaking features and to develop effective countermeasures.

3 Methods

3.1 Definitions

An RDF data set is comprised of statements about certain resources. This statements are encoded in triplets in the form *resource-predicate-object*, where the *resource* is related to any *object* via a *predicate*. The *object* could be another *resource* or a *primitive* (e.g. a double value or a String). An example for such a triplet would be the relation *Austria-export commodity-machinery and equipment*, where the *resource* ”Austria” is related to the *object* ”machinery and equipment” via the *predicate* ”export commodity”. Since, the proposed framework is based on various machine learning techniques, we will further denote the terms *resource*, *predicate* and *object* with the corresponding machine learning terms:

- *Predicate/Feature*: A *resource* is described by various *predicates* (e.g. unemployment rate, gross domestic product). In machine learning these predicates are called *features.*
- *Object/Feature Value*: An object is set into relation with a *resource* by a *predicate.* In RDF such an *object* could be another *resource* or a primitive (e.g. a double value). We will further denote *objects* as *feature values.* Thereby, a given *feature* (e.g. "export commodity") is represented by multiple *feature values* (e.g. "chemicals" or "machinery and equipment").
- *Resource/Instance*: An *instance* represents the collection of all statements made about a given *resource.* For the given data set, the *instance* "Austria" is described by various *features* and their values.

3.2 Related Work

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)][Ding et al.(2005)].

3.3 Employed Machine Learning Methods

Activation Patterns are generated by utilizing three different techniques from the areas of machine learning and artificial intelligence. These building blocks include unsupervised learning algorithms, semantic networks and spreading activation (SA) algorithms. For the analysis and discretization of single features and feature groups we require unsupervised learning algorithms based on prototypes. Examples for such algorithms are Neural Gas (NG) [Martinetz and Schulten(1991)] and its successors Growing Neural Gas, Robust Neural Gas and Robust Growing Neural Gas (RGNG) [Qin and Suganthan(2004)]).

Semantic networks [Quillian(1968)] are directed or undirected graphs that store information in the network nodes and use edges (links) to present the relation between these nodes. Typically, these links are weighted according to a weighting scheme. Spreading activation (SA) algorithms [Crestani(1997)] can be used to extract information from semantic networks. Semantic networks and SA algorithms play an important role within Information Retrieval (IR) systems such as [Fellbaum(1998)][Kozima(1993)][Kozima and Ito(1996)] [Berger et al.(2004)] and [Tsatsaronis et al.(2007)]. By applying SA algorithms we are able to extract *Activation Patterns* from trained semantic networks. These *Activation Patterns* can then be analyzed by arbitrary unsupervised learning algorithms such as Self Organizing Maps (SOM) [Kohonen(1995)], Hierarchical Agglomerative Clustering (HAC), Expectation Maximization (EM), k-means, etc.

3.4 From RDF triplets to *Activation Pattern*

The process of generating and analyzing *Activation Patterns* is separated into five processing layers (L1-L5) depicted in Figure 1. The general idea is to extract the co-occurence information of different features (L1, L2), to store this information in an semantic network (L3) and to generate *Activation Patterns* by applying spreading activation (SA) strategies (L4). Various analysis techniques can then be applied to the generated patterns (L5).

3.4.1 L1 - Feature extraction

This layer extracts instances and their features from a given RDF data set. In addition, arbitrary instance[5] and feature filters[6] can be applied according to the L5 analysis requirements. Regardless of the nature of the analyzed RDF data set, the extracted features of any data set can always be separated into these two categories:

- **Distance-based features:** The feature values of such features can be put into relation by defining a distance-measure such as the Euclidean distance. An example is the unemployment rate in percent: We can clearly say that an unemployment rate of 5% is closer to 10% than it is to 30%.
- **Symbolic features:** For these features we cannot find a meaningful distance relation between the feature values. An example is the country name: The feature value *Austria* is not closer to the feature value *Germany* than it is to *Italy*.

Feature extraction - summary

1. A given RDF data set is parsed and the instances for a given resource-set are extracted (e.g. all persons and their features from a corporate database).

[5] e.g. only African countries

[6] e.g. only feature related to exports

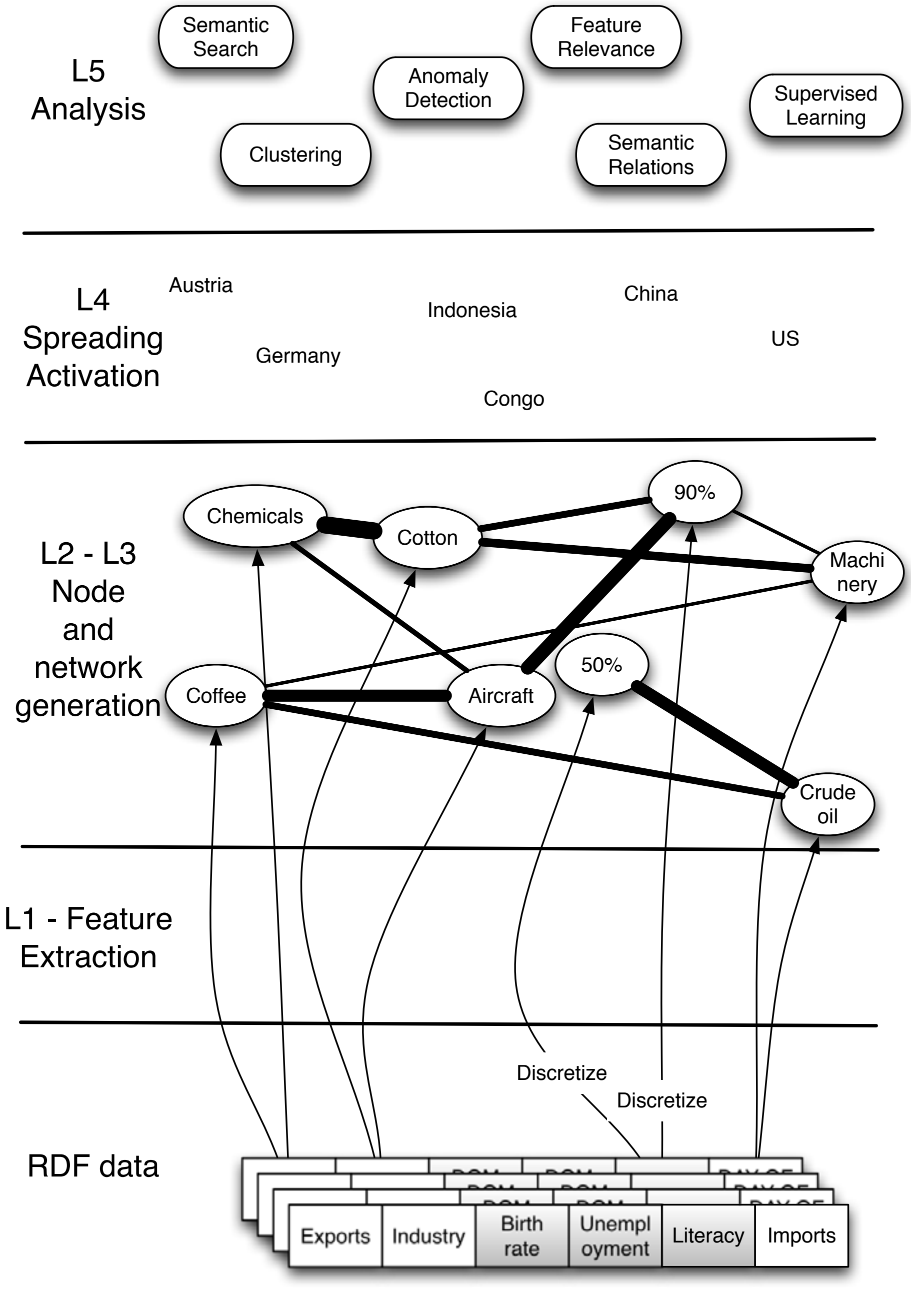


Figure 1: Processing layers for the Activation Pattern transformation and analysis

2. The features of the extracted instances are filtered according to the instance and feature filters.

3. The features are tagged according to their type: *distance-based* or *symbolic*.

3.4.2 L2 - Node generation

This process layer creates the nodes for the semantic network generation process in L3. The process of mapping feature values to nodes depends on the type of the particular feature. For *symbolic* features the possible values are directly mapped to separate nodes. For *distance-based* features we need to apply some kind of discretization operation to map values onto nodes.

Although there is a wide range of discretization algorithms available, we have chosen the RGNG algorithm. It is applied to the real feature values of a *distance-based* feature and the trained prototypes are used as nodes for the semantic network. Basically any prototype-based unsupervised learning algorithm could be used for the discretization process. However, RGNG was selected, since it includes several robust learning techniques and employs the Minimum Description Length (MDL) [Rissanen(1989)] to automatically determine the model complexity. Since the performance of RGNG and similar algorithms has been evaluated by applying them to a wide range of data sets, we can assume that these algorithms will produce good results for the low dimensional data represented by single features. Although the computational complexity of RGNG is high, the benefits justify its application and improve the employed analysis techniques. In other more specific scenarios, the RGNG algorithm can be replaced with a simple adequate discretization method.

Figure 2 shows a simple example where a *distance-based* feature is mapped to semantic network nodes: In this case a RGNG map with 4 prototypes (clusters) was trained on the values of a percentage-based feature. A node is generated for each of these prototypes and the feature values within the analyzed instances are then mapped to these nodes. The value range covered by each prototype depends on the location of input value agglomerations (clusters).

Node generation - summary

1. All the values of *symbolic-features* are directly mapped to unique nodes within the semantic network. E.g. for the feature *border-country* the nodes *border-country:Austria*, *border-country:Germany* etc. are generated.

2. For each *distance-based* feature, an RGNG-map is trained for the corresponding feature values. The prototypes of the trained RGNG-map are directly mapped to unique nodes within the semantic network.

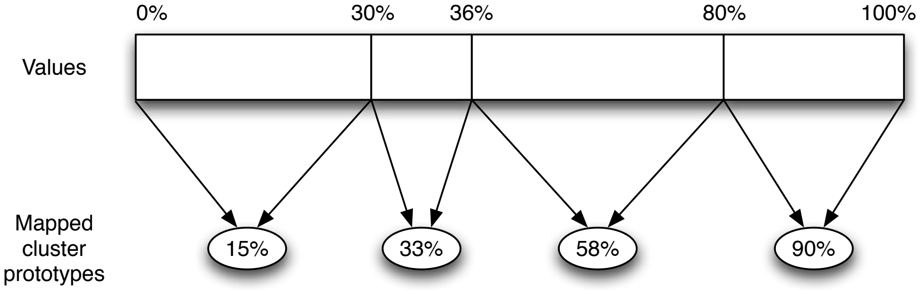


Figure 2: Mapping of values to prototypes for a *distance-based* feature based on percentage values.

3.4.3 L3 - Network generation

This layer analyzes the relations between the feature values that co-occur within the instances. The nodes within the network represent the feature values and the relations are modeled by creating links between the network nodes. The strength of these links depend on how often feature values co-occur. Again, the mapping of the feature values to the network nodes depends on the nature of the features: *Symbolic* feature values are mapped directly, whereas distance-based feature values are mapped via the RGNG prototypes, trained in L2.

The link weights of the generated network represent the absolute number of co-occurence between the different feature values. In order to apply the spreading activation function in L3, these weights must be normed. During the evaluation of the framework we focused on a local norm, that norms the outgoing links of a given node according to the maximum link weight of these links. This local norm emphasizes feature values that sparsely occur within the data set, which allows for a better analysis of their relations.

The complete network generation process is depicted in Figure 5. The example shows various country instances with different features and the corresponding semantic network.

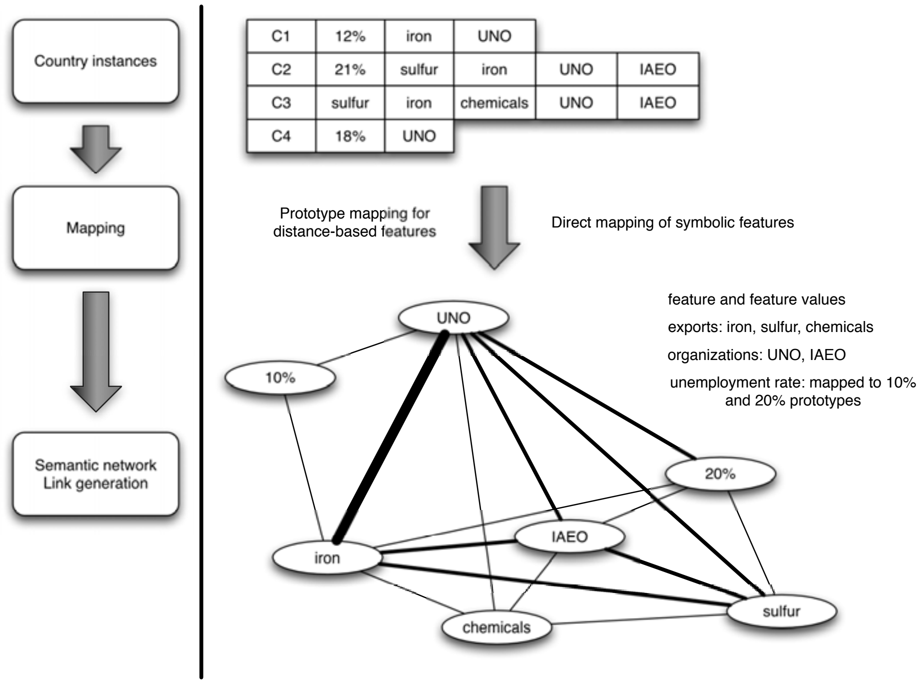
Network generation - summary

1. For each instance we apply the next two steps:

2. The feature values of the given instance are mapped to nodes within the semantic network. *Symbolic* feature values are directly mapped. *Distance-based* feature values are mapped according to the RGNG prototypes. The corresponding nodes are extracted.

3. Links are generated between each of these nodes. Thereby, newly created links between two nodes are initialized with weight 1. The weight of existing links is increased by 1.

4. After processing all instances the links of the generated network are normed according to a local max norm.



C1	12%	iron	UNO		
C2	21%	sulfur	iron	UNO	IAEO
C3	sulfur	iron	chemicals	UNO	IAEO
C4	18%	UNO			

Country instances
Mapping
Semantic network
Link generation
Prototype mapping for distance-based features
Direct mapping of symbolic features
feature and feature values
exports: iron, sulfur, chemicals
organizations: UNO, IAEO
unemployment rate: mapped to 10% and 20% prototypes
UNO
10%
20%
IAEO
iron
chemicals
sulfur

Figure 3: A simple example for an semantic network that was trained on country instances: The strength of the links depends on the strength of the relation between the feature values.

3.4.4 L4 - *Activation Pattern* Generation

This layer applies spreading activation (SA) techniques to the semantic network in order to extract information about features, feature values or instances. The SA technique is a simple process based on these steps:

1. Select one or more feature values and activate the corresponding nodes within the semantic network. In case of a given instance, the nodes for all feature values are activated.

2. For each activated node spread the activation over the links emanating from the node to all neighboring nodes. Given a source node with the activation value a_i, a link emanating from the source node with the weight w and a decay factor D, the activation energy a_j of the neighboring node can be calculated in this way: $a_j = a_i w D$ where D is a value between 0 and 1. A reasonable choice for D is a value around 0.3 since it ensures that neighboring nodes have an adequate influence.

3. Extract the activation value of all nodes within the network and store them in a vector.

We call the vector generated in step 3 *Activation Pattern.* It represents the activation state of the network and varies according to the regions activated within the network. Therefore the *Activation Patterns* allow us to compare different network states with a simple similarity measure such as the Cosine similarity[7]. Figure 4 depicts two examples for *Activation Patterns.* The differences in the patterns indicate that different regions of the network have been activated.

The selection of the nodes to be activated depends on which information we want to extract but is typically based on these scenarios:

- **Activation of chosen nodes:** In this case just a new nodes corresponding to *feature values* are activated. This scenario is used when we need to extract information about the relations between features.

- **Activation of complete *instances*:** The standard L4 process includes the generation of *Activation Patterns* for all instances. For each instance, all of its feature values are extracted, the corresponding nodes in the network are activated, the activation is spread and the *Activation Pattern* is extracted from the activated network.

[7] The Cosine similarity is better suited to model the similarity between activated networks than the Euclidean distance: Two *Activation Patterns* that activate distinct regions within the network are orthogonal – meaning not related – and their distance would be $\frac{\pi}{2}$. This orthogonality and therefore non-similarity of patterns could not be modeled with the Euclidean distance.

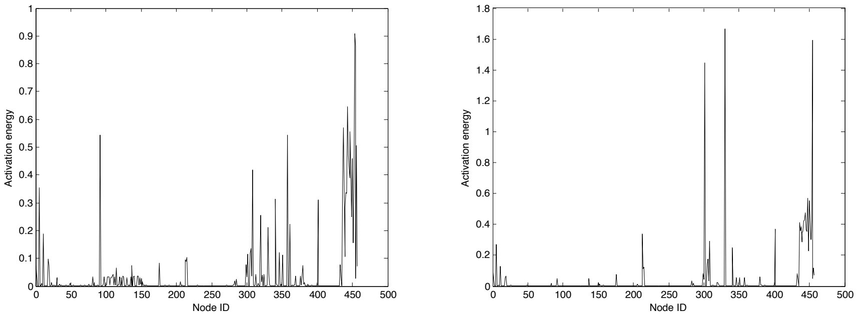


Figure 4: Examples for two different *Activation Patterns*, the x-axis represents the nodes within the network, the y-axis represents the activation energy of these nodes after applying the spreading activation process to the activated nodes.

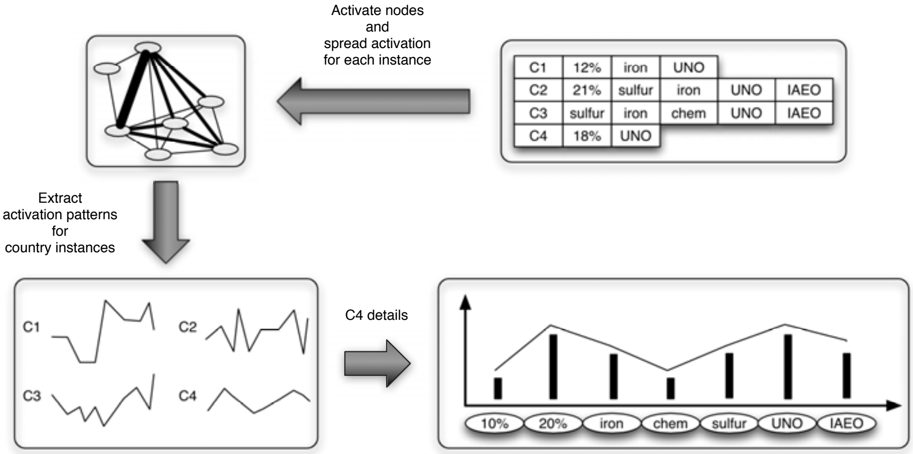
***Activation Pattern* generation - summary**

1. For each instance execute the next two steps:
2. Activate all nodes within the network that correspond to the feature values within the instance.
3. Spread the activation of these nodes to the neighboring nodes according to the SA algorithm.
4. Extract the *Activation Pattern* from the network.

3.4.5 L5 - Analysis

The *Activation Patterns* generated in L4 are the basis for all further analysis procedures:

- **Unsupervised clustering:** Due to the combination of arbitrary features with different value ranges and different meanings, a normalization strategy is required for a typical unsupervised learning scenario. Since the *Activation Pattern* transformation analyzes the relations between different features and not their values, normalization is not required. The transformation has the additional advantage that we can easily combine *distance-based* and *symbolic* features, while being able to apply standard *distance-based* unsupervised clustering algorithms to the generated *Activation Patterns* without losing information about the semantic relations within the data set. By varying the model complexity of the employed unsupervised learning algorithm, we are able to build a hierarchy from a very coarse grained categorization down to a very detailed representation of the analyzed data.



C1	12%	iron	UNO		
C2	21%	sulfur	iron	UNO	IAEO
C3	sulfur	iron	chem	UNO	IAEO
C4	18%	UNO			

Figure 5: Spreading activation and pattern modeling: First, the nodes corresponding to a given instance (e.g. C4) are activated. Second, the activation is spread over the network via a spreading activation technique. Finally, the *Activation Pattern* is generated by extracting the activation values of all network nodes and arranging them in a vector (e.g. the C4 *Activation Pattern* in the last Figure).

- **Semantic search:** The distance between the *Activation Patterns* can be used to implement semantic search algorithms that retrieve semantically related instances. These search queries can also be used to specify certain feature values and find closely related patterns.

- **Feature relations:** The semantic network describes arbitrary relations between feature values. By activating one or more nodes (corresponding to feature values) within the semantic network, and spreading their activation via the links to the neighbors, we are able to extract details about the relations between various feature values and the strength of these relations.

- **Feature relevance:** The relations within the semantic network are created according to the co-occurrence of feature values within the analyzed data set. The strength of these relations are represented by the associated weights within the network. Given a feature value that is represented by a node and the number of emerging/incoming links and their weights, we are able to deduce the importance of the information carried by the node. Nodes that are connected to a large number of other nodes typically do not add information for subsequent analysis processes. This is highlighted by a simple example:

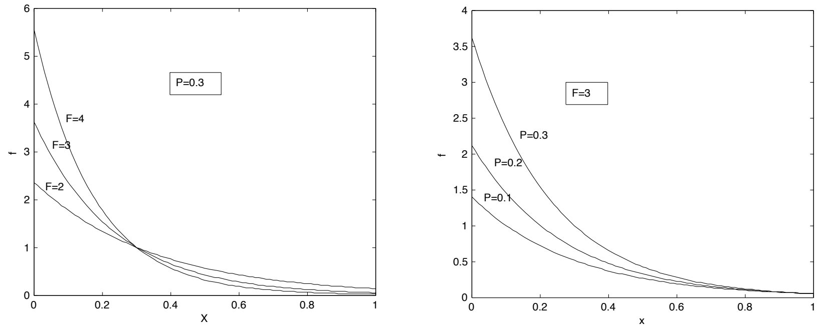
Assuming we analyze english speaking countries, the node for the feature value *language:english* does not carry any information at all since all of the analyzed countries have the same language. The node is therefore connected to all feature values of the analyzed instances. In order to penalize such nodes and thereby attenuate the influence when applying SA, we calculate the fanout values for the each node:

Given a semantic network with n nodes, the maximum link weight 1 and an arbitrary number of links with different weights l_{ij} between nodes i and j:

$$\begin{aligned} x_i &= \frac{\sum_{j=1}^{n} l_{ji}}{n-1} \\ f_i &= \frac{e^{(1-(x_i-P)(\frac{F}{1-P}))}}{e} \end{aligned} \tag{1}$$

The value x_i for the node i is the sum of all weights from incoming links l_{ji} divided by the number of maximum links per node $n-1$.

The fanout value f_i for node i depends on F and P. F is used to shape the fanout function and P is the percentage of the maximum possible sum of incoming link weights for a given node. Due to the max norm that is applied to all emerging links the maximum sum of all incoming links to a given node i is equal to $n-1$. This happens when the feature value represented by node i is equally connected to all other values. If a node has no links at all then the link sum is equal to zero. P is used to specify the maximum sum of incoming link weights that does not cause the attenuation of the node activation value. The influence of the parameters P and F are shown in Figure 6.



(a) Fanout functions for $P = 0.3$ and $F = 2$, $F = 3$, $F = 4$

(b) Fanout functions for $F = 3$ and $P = 0.1$, $P = 0.2$, $P = 0.3$

Figure 6: Influence of the parameters F and P

4 CIA World Factbook RDF analysis

In order to show the benefits of the *Activation Patterns* concept we have analyzed an RDF representation[8] of the *CIA World Factbook*[9]. Since the features used to describe the countries are well known, this RDF data set is a perfect choice for demonstrating and evaluating the *Activation Pattern* concept. The CIA Worldfactbook describes all countries by a large number of *symbolic* and *distance-based* features. During the feature extraction process we have filtered several features that do not provide interesting information for the further analysis processes. These include features that have a textual description as value (e.g. background information about a country), features that are unique for each country (e.g. name of lowest and highest elevation point) or certain features that use absolute numbers (e.g. number of televisions within a country)[10]. However, not all of these low-information features were removed in order to show that the feature relevance calculation works appropriately.

All of the following procedures were executed on a Java framework that implements the layered structure of the *Activation Pattern* concept[11].

1. The CIA Worldfactbook RDF data set was parsed with the JENA framework[12] and imported into the analysis framework.

2. For each described country an instance was generated including all the features that were not filtered (L1 Section 3.4.1).

3. The features were tagged according to their type: *distance-based* or *symbolic* (L1 Section 3.4.1).

4. The semantic network was generating by creating a node for each feature value. For the *distance-based* features RGNG maps were trained (the maximum number of prototypes was limited to 10) (L2 Section 3.4.2).

5. The relations between the feature values of the country instances were mapped to the links within the network. The links of each nodes where normed with a local max norm (L3 Section 3.4.3).

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

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

[10] The *Activation Pattern* concept does not require normalization of features with different value ranges. However, in this case the problem is related to different feature representations. E.g. the absolute number of televisions does not carry valuable information without putting it into relation with another feature (e.g. the whole population). We could have normed this and other features with the absolute population of a country, however it was not necessary since there was still valuable information in other features.

[11] The development of the framework is under progress. The framework will be released as open source project in the near future.

[12] http://jena.sourceforge.net/

6. The *Activation Patterns* where then generated by applying spreading activation to the network which was activated by the feature values of the instances. (L4 Section 3.4.4).

7. The analysis procedures of L5 (Section 3.4.5) were applied to the *Activation Patterns*. The results are described in the remainder of this section.

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)
Relation 3	*agricultureProduct: bananas*		*agricultureProduct: tomatoes*	
unemploym.	24.45 (1.0)	12.93 (0.8)	12.93 (1.0)	04.02 (0.6)
literacyTotal	95.92 (1.0)	80.06 (0.5)	95.92 (1.0)	80.06 (0.3)
grossAgriculture	14.80 (1.0)	03.68 (0.5)	03.68 (1.0)	14.80 (0.6)
grossServices	60.15 (1.0)	70.37 (0.6)	70.37 (1.0)	60.15 (0.6)
grossIndustry	21.24 (1.0)	30.39 (0.5)	10.38 (1.0)	30.39 (0.8)
exports	coffee (1.0)	bananas (0.9)	textiles (1.0)	agricultural products (0.7)

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.

Feature	**Type**	**Description**
industry	sym	e.g. food processing, auto parts, chemicals
exportsCommodity	sym	e.g. diamonds, metal goods, livestock
language	sym	e.g. English, German
naturalResource	sym	e.g. wildlife, copper, salt, fish, oil
importsCommodity	sym	Same values as exportCommodity
agricultureProduct	sym	e.g. coffee, cocoa, coconuts
border-country	sym	Same values as desccountry
exportPartner-country	sym	Same values as desccountry
participatesIn	sym	e.g. G-24, IAEA, UNO
importPartner-country	sym	Same values as desccountry
environmentalAgreement	sym	e.g. Antarctic Treaty, Biodiversity
mapReferences	sym	e.g. Europe, Asia
highestPoint-elevation	real	Altitude (meters), absolute value
literacyTotal	real	Literacy rate, percent
unemploymentRate	real	Unemployment, percent
airports	real	Number of airports, absolute value
latitude	real	Latitude, absolute value
grossDomesticProduct	real	$, absolute value
highwaysUnpaved	real	Length in km
malesFitForMilitaryService	real	Absolute value
mainTelephoneLines	real	Absolute value
heliports	real	Absolute value

Table 2: Description of features and their types used for calculating the relevance values.

4.1 Relations between Objects

The generated semantic network stores knowledge about the semantic relations between features and feature values. By activating nodes within the network, spreading the activation to neighboring nodes we are able to extract these semantic information:

1. Select a feature value and activate the corresponding node within the semantic network by setting the activation value to 1.0.

2. Spread the activation to semantically related nodes by applying the spreading activation technique described in L4 (Section 3.4.4).

3. Extract the *Activation Pattern* by storing all node activation values in a

Feature	**Relevance**	**Feature**	**Relevance**
industry	1124.5	environmentalAgreement	36
exportsCommodity	875.5	mapReferences	25.5
language	732.6	highestPoint	21.1
naturalResource	512.8	airports	12.5
importsCommodity	479.4	latitude	11.5
agricultureProduct	399.5	grossDomesticProduct	8.3
border-country	309.6	highwaysUnpaved	8.3
exportPartner	234.4	malesFitForMilitaryService	8.1
participatesIn	198.3	mainTelephoneLines	6.6
importPartner	179	heliports	4.3

Table 3: Feature relevance of selected features. A higher value indicates a large influence.

vector.

4. Analyze the pattern: In order to highlight the strength of the relation for each feature, we norm the activation values (corresponding to the feature values) by their maximum value.

Table 1 shows the normed activation values for selected features when executing the following queries:

- **Relation 1 -** *"How do the typical values for unemployment rate, literacy and gross domestic product sectors and exports compare between Africa and Europe?"*

  By activating the node for the feature value *Africa* of the feature *mapReference*, 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*.

  In contrast, columns 3 and 4 show the results for countries in *Europe*, which are – as expected – typical for richer countries. This is indicated by a high literacy rate, low unemployment, a large service sector and only a very small agriculture sector. Furthermore, the country exports are not based on raw materials (e.g. coffee) but on industrial goods.

- **Relation 2 -** *"How do the same values compare for countries that export crude oil vs. countries that export machinery, equipment and chemicals?"*

The results for crude oil exporters (columns 1 and 2) indicate a larger unemployment rate, a larger industrial sector and a smaller service sector when compared to countries that export machinery, equipment and chemicals (columns 3 and 4).

- **Relation 3 -** *"How do the typical values for unemployment rate, literacy, gross domestic product sectors and exports compare between countries that export bananas and countries that export tomatoes?"*

These three examples show, that by activating one single node within the network we are able to extract valuable information about the semantic relations between different features and their values. The results of these two example relations are not surprising, however such well known relations were used to verify the functionality of the *Activation Pattern* transformation. The same process can now be used to discover semantic relations within other data sets where we do not have background information about the features and their relations.

4.2 Feature relevance

By using the fanout information generated in L3, we are able to attenuate the impact of nodes that do not carry information due to their high interconnection rate to other nodes. For the CIA world factbook a good example would be the feature *population*, which represents the absolute number of people living within a given country. Since the population size of a country is not related to other features such as economic power, unemployment rate, exports, imports etc.[13] the feature co-occurs randomly with other features and is connected to a large number of nodes within the network.

In order to determine the relevance of the analyzed features we extract the fanout values of each value and sum them up for each feature. The results for the most and least important features and the descriptions of the features are shown in Table 3 and in Table 2.

4.3 Semantic Aware Search Queries

The similarity of generated *Activation Patterns* – and therefore their semantic relatedness – can be calculated by comparing *Activation Patterns* with a distance-measure (e.g. cosine similarity). This enables us to find patterns that activate similar regions on the semantic network and are therefore related. By selecting an existing pattern (e.g. for *Austria*), similar countries can be retrieved

[13] Although the population size is used to derive other features such as unemployment rate etc., the value alone does not provide any information in respect to these features. The unemployment rate of a small country can be as high/low as that of a large country.

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 4: Results for a search query that retrieves countries that are semantically related to crude oil exporting countries, but do not have the feature value *crude oil*.

by comparing the *Activation Patterns*. 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.

- **Query 1 -** *"List all countries according to their similarity with Austria"*

  The *Activation Pattern* for Austria is taken and compared to the patterns of all other countries by utilizing the cosine similarity as distance measure. The best matching countries are Germany, Sweden, Switzerland, Netherlands and the most unrelated are Sudan, Gaza Strip and the West Bank.

- **Query 2 -** *"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"*

  In this case, we activate the *crude oil* node, generate the *Activation Pattern* and search for similar country patterns. The results for the 11 best matching countries are not shown, since they are *crude oil* exporters and 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*[14] 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*.

[14] 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.

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 5: Examples for the strength of semantic relations between various feature values.

4.4 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 5 only the features 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.

5 Conclusions and Future Work

In this article we demonstrate the benefits of utilizing *Activation Patterns* for the analysis of RDF data sets that typically consists of a wide range of *symbolic* and *distance-based* features. Due to the transformation process that focuses on the relations between the features and not theirs values, the *Activation Patterns* form the basis for a wide range of semantic analysis techniques based on machine learning procedures.

We have already successfully applied the concept to a wide range of data sets. Future work will make refinements to the implemented algorithms and the focus will be placed on the completion and improvement of the existing Java framework. This should allow us to easily extend the concept to arbitrary other domains such as the temporal and semantic analysis of data that evolves over time.

References

[Berger et al.(2004)] Berger, H., Dittenbach, M., Merkl, D.: “An adaptive information retrieval system based on associative networks”; (2004).

[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.

[Kohonen(1995)] Kohonen, T.: Self-Organizing Maps; volume 30 of Springer Series in Information Sciences; Springer, Berlin, Germany, 1995.

[Kozima(1993)] Kozima, H.: “Similarity between words computed by spreading activation on an english dictionary”; EACL; 232–239; 1993.

[Kozima and Ito(1996)] Kozima, H., Ito, A.: “Context-sensitive measurement of word distance by adaptive scaling of a semantic space”; volume cmp-lg/9601007; 1996.

[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).

[Rissanen(1989)] Rissanen, J.: Stochastic Complexity in Statistical Inquiry Theory; World Scientific Publishing Co., Inc., River Edge, NJ, USA, 1989.

[Teufl and Lackner(2010)] Teufl, P., Lackner, G.: “Rdf data analysis with activation patterns”; K. T. und Hermann Maurer, ed., 10th International Conference on Knowledge Management and Knowledge Technologies 13 September 2010, Messe Congress Graz, Austria; Journal of Computer Science; 18 – 18; 2010.

[Teufl et al.(2010a)] Teufl, P., Lackner, G., Payer, U.: “From nlp (natural language processing) to mlp (machine language processing)”; I. V. Kotenko, V. A. Skormin, eds., Computer Network Security, 5th International Conference on Mathematical Methods, Models and Architectures for Computer Network Security, MMM-ACNS 2010, St. Petersburg, Russia, September 8-10, 2010, Proceedings; volume 6258 of Lecture Notes in Computer Science; 256 – 269; Springer, 2010a.

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