Master Thesis

The application of self-organizing maps on semantic data in the form of a topic map

Q.H.J.F. Siebers BSc.

Master Thesis DKE 12-28

Thesis submitted in partial fulfillment
of the requirements for the degree of Master of Science
of Artificial Intelligence at the Department of Knowledge
Engineering of the Maastricht University

Thesis Committee:

Prof. Dr. G. Weiss
Dr. M. Winands

Maastricht University
Faculty of Humanities and Sciences
Department of Knowledge Engineering
Master Artificial Intelligence

August 31, 2012

Preface

This paper is my master thesis. It describes the research I performed while enrolled at the Department of Knowledge Engineering at Maastricht University. In this thesis I propose two methods of vectorization of Topic Maps data and apply these methods on a static SOM and a GHSOM network. I would like to thank my supervisor Prof. Dr. Gerhard Weiss for the continued support during the long road that lead to this thesis. I would like to thank Dr. Mark Winands for the reviewing my thesis and providing corrections that improved this thesis. Finally, I would like to thank Peter-Paul Kruijsen and Gabriel Hopmans of Morpheus Kennistechnologie for their support and providing the time for me to finish this thesis.

Quintin Siebers
Weert, August 2012

Abstract

Self organizing maps (SOMs) is a popular technique for clustering data in the text mining field. Topic Maps is a standard for storing knowledge comparable to RDF. Unlike a document made up of words that indicate semantic value, a topic map stores this semantic value in its different constructs. The application of SOMs on Topic Maps has not yet been researched.

This thesis researches the application of SOM networks on Topic Maps based data. Four research questions are used to answer this problem statement. The first two of these questions focus on the vectorization process, the other two on different SOM techniques.

To test the proposed vectorization methods, a corpus is needed that contains enough data to create a large topic map. The Internet Movie Database (IMDb) provides it's core data for non commercial use. A part of this corpus was converted to a topic map and split into ten test topic maps.

This thesis proposes two methods of topic map data vectorization. The first is a method based on counting the words in the textual data in the topic map. The second method is based on counting ontological properties of topics.

The ten test data sets are vectorized with the two proposed vectorization methods, and used to train a static SOM and a Growing Hierarchical SOM (GHSOM). The resulting SOM networks show that all combinations of both vectorization methods and both SOM techniques result in a clear clustering of topic map based data. The data based vectorization in combination with the GHSOM provided the most detailed result, the ontology based vectorization in combination with the static SOM the most abstraction.

Based on these results the research questions and problem statement were answered. The proposed methods and tested techniques can be used for clustering topic maps. The choice which method and technique to use depends on the required level of abstraction needed.

Contents

Chapter 1

Introduction

Most applications of self-organizing maps (SOMs) are focused on clustering a document space in the field of text mining. Documents are clustered based on the frequency of words appearing in them. In some cases meta-data is used to enrich the document vector. This application of SOMs has been proven to work [11].

With the development of more semantically rich environments and the introduction of the 'Big Data' research field, more large scale and semantically valued data analyzing is required. Semantic value rich data can be stored and manipulated in several ways, one of which is the Topic Maps standard.

This chapter provides an introduction into the techniques used within this thesis. It also provides the problem statement and research questions.

1.1 Topic Maps

Topic Maps is an ISO standard that allows for the modeling and representation of knowledge in an interchangeable form, and provides a unifying framework for knowledge and information management [4]. Topic Maps is based upon the characteristics of the index of a book, linking *topics* to *sources* by using page numbers (*occurrences*). It extends the power and ideas of SGML, XML, semantic networks, RDF and Frames by adding more complex but intuitive solutions and possibilities [7, 16, 17, 14]. Each of the objects in a topic map[1] has a specific semantic value [5].

Topic Maps is closely related to the semantic web because Topic Maps is interchangeable with RDF. A website driven by Topic Maps can easily be extended to be part of the semantic web by hosting RDF fragments on each page.

[1] The term 'Topic Maps' refers to the ISO standard; the term 'topic map' refers to an instance of the standard.

1.1.1 Topics

A topic is a representation of a subject in any domain. This subject can be anything whatsoever. There is just one basic constraint on a topic: one topic per subject. Topics are the main focus of this thesis. They will be used as input documents for the SOM.

1.1.2 Names

Because a topic represents a subject and not the name of a subject (which can differ in context), it can have multiple names. A name has a specific type, which adds to the semantic value of the name.

1.1.3 Occurrences

Topics can have occurrences. These occurrences store non-relational or external information links. Like names, occurrences has a type which adds to the semantic value of the occurrence. Examples of occurrences are a date of birth value or a reference to a website.

1.1.4 Associations

Associations contain the largest amount of semantic value within a topic map. As their name suggests, they associate between topics, creating a semantic web of topics. To add semantic direction associations are made up of roles. An association must have at least one role. Associations and roles have types which increase the semantic value of the relationship.

1.1.5 Scope

Names, occurrences and associations can be given a scope value to specify a context. From [3]:

> "Scope is a feature of Topic Maps used to represent the qualification of a statement. That is, the scope represents the context of validity for the statement. Any statement in Topic Maps is qualified using a set of topics, possibly empty, and this is called the scope of the statement. The empty scope is known as the unconstrained scope, and statements in this scope are considered to have unlimited validity within the context of a particular topic map."

Therefore, the presence — or absence — of scope on a characteristic of a topic gives that characteristic more semantic value.

1.1.6 Reification

Any non-topic object within a topic map can be reified. *Reification* is the process of extending with a topic. For example: reifying an employment association to

register starting date and a contract for the employment. The presence of reification on a characteristic increases that characteristic's semantic value.

1.2 Self Organizing Maps

A self-organizing map (SOM) is a type of artificial neural network that is trained using unsupervised learning to produce a low-dimensional, discrete representation of the input space of the training samples. This representation is called a map because it is typically two-dimensional. Self-organizing maps are different from other artificial neural networks in the sense that they use a neighborhood function to preserve the topological properties of the input space [12].

1.2.1 Growing Hierarchical SOM

The Growing Hierarchical SOM (GHSOM) was introduced in [2] as

> "... a dynamically growing neural network model evolves into a hierarchical structure according to the requirements of the input data during an unsupervised training process."

Its main strength is the ability to find and map hierarchical structures in the presented document space. This makes it a good candidate for mapping semantic data, because it can better distinguish between inputs with a large amount of equal feature values. For example: the fact that a topic has a name can be seen by the SOM as a principal feature. A GHSOM will further expand this cluster to detect underlying clusters.

1.3 Vectorization

In order for the Self Organizing Map (SOM) to be able to work with the data stored in a topic map it first needs to converted into a set of vectors. This process of converting is called *vectorization*. The expected output of vectorization is a vector for each document in the input data. Each vector within the set is expected to have the same size.

The commonly used approach for vectorization is to count the appearances of words in the documents and form vectors (per document) from the counts of the words in the document space [12]. The SOM is able to recognize, and preserve, the semantics of sentences. However, within a topic map different types of characteristics are not always represented by a piece of text. Furthermore, there are no sentences that hold semantic value because the constructs of a topic map imply these semantics.

1.4 Problem statement

A research into the combination of Topic Maps data and SOM techniques is needed. This thesis is an initial research into this combination. The problem

statement of this thesis is:

Problem statement How can existing SOM techniques be applied to semantic data represented in a topic map?

Because SOM techniques work on a set of mathematical vectors, a vectorization is needed. This point leads to the following four research questions:

Research question 1 Can a vectorization method be defined that uses the textual values stored in a topic map? A possible problem is the the low amount of textual data in the topic map.

Research question 2 Can a vectorization method be defined that uses the ontological properties of a topic? The idea is to count these properties instead of counting the words connected to the topic.

Different SOM techniques might provide different results. Therefore, the following two research questions are related to testing different SOM techniques:

Research question 3 Can the static SOM technique be applied to the two proposed vectorization methods?

Research question 4 Can the GHSOM technique be applied to the two proposed vectorization methods? This might give more insight into the structure of the data compared to a static SOM.

1.5 Thesis layout

To test the proposed methods a corpus topic map is created and cut into smaller test sets, as described in Chapter 2; 'Corpus'.

The two methods of vectorization of topics in a topic map are described in Chapter 3; 'Vectorization'.

Chapter 4, 'Experiments', describes how the two methods are compared using the test sets. The results of these tests are presented, per technique, in the 'Results' section.

Finally, Chapter 5, 'Conclusions', sums up the results and provides a conclusion whether Topic Maps data can be clustered using SOM techniques in combination with the two proposed vectorization methods.

Chapter 2

Corpus

A well defined, structured and large data set is needed to create a good test set for the proposed methods. This is called a corpus.

This chapter describes the corpus used, and the transformations needed, for the research in this thesis.

2.1 The Internet Movie Database

The Internet Movie Database (IMDb) is an online record of movies, series, games, television productions and actors. IMDb started in 1990 as a follow up of a newsgroup that kept track of cinematographic facts. The database contains about 2.2 million titles and 2.4 million actors [8, 10, 15].

A large part of the data set is available for noncommercial use in the form of plain text lists [9]. Each list contains a certain section of the facts collected by IMDb. For example, there is a file for the genres of each title.

The IMDb data set is well defined and structured, which helps analyze the results of the clustering discovered by a SOM method. The large size of the data set makes it possible to select several sub selections of data to test without having to use the same data over and over.

2.2 Size reduction

As mentioned in the introduction, the IMDb data set is quite large. The corpus is reduced in size due to the fact that the research will only focus on selections of the complete data set. This reduction is based on several selection criteria.

2.2.1 Titles

Titles must have been created within the time span of 1990 to 2012. The reason for this criteria is that the IMDb only started in 1990 and the titles before 1990 are not fully represented and likely incomplete. The other side of the range was

chosen because any title registered in the future is also often incomplete. The remaining set of titles is approximately 64% of the complete set of titles.

Note that this criterion also filters out any titles with an unknown creation date.

2.2.2 Actors and actresses

The set of actors and actresses is the biggest in the IMDb corpus. The title selection criteria already limits the number of possible connections, but not enough to make the conversion of the corpus executable in an acceptable time span. Therefore, all actors that are included in the selected corpus have more than one title to their name, counted over the actually present titles, and less than 200 titles. These actors and actresses are a nice addition to our selected data set because they are not just loose facts about a title, but actually create a network of interconnected topics.

This selection criteria results in a selection of approximately 33% of all actors and actresses from the complete corpus.

2.2.3 Inter title relations

Another network-creating characteristic provided in the IMDb corpus is the list of inter title relations. These relations indicate references, spoofs, versions, sequels, remakes, features, spin-offs and edits between titles. Any connection between titles that passed the previous selection criteria is included in the selection.

2.2.4 Other included characteristics

Beyond the titles, actors and inter title relations, the selection also includes:

- Genre
- Country of production
- Language
- Filming location
- IMDb user rating

Only characteristics for titles that passed previous selection criteria are included in the selection.

2.3 Resulting data set

The resulting topic map contains 4.7 million topics, 516 thousand occurrences and 10.1 million associations. Figure 2.1 depicts the distribution of topic types in the selected data set.

The reason for the huge number of character instances is that there is no way of telling if two characters with equal name are in fact the same character when viewed over several titles. Characters were merged whenever this information was derivable (due to the same series, for example).

The other previously unmentioned types are:

Episode A TV series is made up of several episodes.

Note Note topics are used to scope character role associations to indicate additional properties. For example, if an actor only acts out the voice of a character the note for this association will be named *"voice"*.

Video The video type stands for titles that are produced for video distribution only.

TV The TV type stands for titles that are produced for television only.

Series The series type represents a collection of episodes with a collective name.

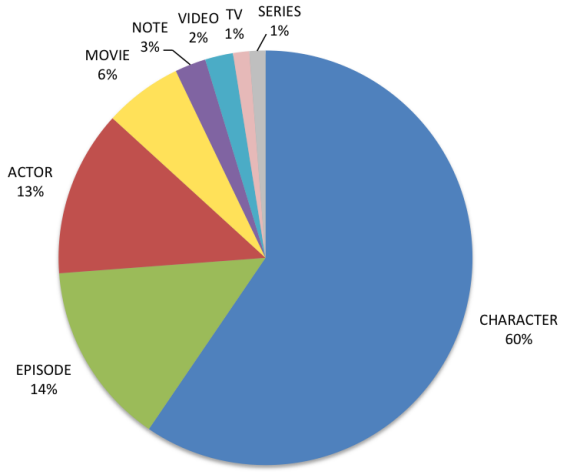


Figure 2.1: Distribution of topic types in the selected data set

2.4 Selection

From this complete topic map several smaller topic maps were created by randomly adding topics from the complete test topic map and their direct connected topics and characteristics. The topics that were selected in this manner are given a flag type upon processing so the vectorization methods can recognize them.

Table 2.1 shows the total item counts for each of the created test sets. It also shows the type counts for the focus topics which shows how representative

Test set	0	1	2	3	4
Total topics	7081	7619	7697	8024	7457
Total associations	4264	4610	4626	5011	4379
Total occurrences	105	118	103	108	104
Character	627	576	606	594	589
Episode	122	153	131	138	150
Actor	120	133	136	142	127
Movie	59	66	56	62	50
Note	26	20	24	18	30
Video	27	23	20	21	22
TV	16	14	16	11	15
Series	3	15	11	14	17

Test set	5	6	7	8	9
Total topics	7838	8210	8133	8655	6729
Total associations	4831	4889	4881	5381	4068
Total occurrences	99	102	89	119	102
Character	634	603	601	602	619
Episode	128	136	140	141	130
Actor	125	142	147	125	128
Movie	57	59	45	60	53
Note	14	17	22	13	21
Video	16	24	22	27	22
TV	12	9	12	15	12
Series	14	10	10	17	15

Table 2.1: Test set statistics and type counts of selected topics

the topic maps are for the complete corpus. The object count statistics are similar for all of the test sets. Therefore, the connectivity (the average number of association per topic) of the test sets approximates the connectivity of the complete data set. Although topics of supporting types — like language, genre, location and country — are not selected as focus topics they are still included in the topic maps because of references to them from the selected topics.

Chapter 3

Vectorization

As explained in the introduction, vectorization is the process of converting the input data into a set of vectors. This chapter describes two proposed methods of vectorization: As explained in the introduction, vectorization is the process of converting the input data into a set of vectors. This chapter describes two proposed methods of vectorization:

1. The data based transformation
2. The ontology based transformation

Due to the way the complete data set was divided into smaller topic maps the vectorization algorithms will only convert the topics that were focus points during the selection process as described in Section 2.4.

3.1 Data based vectorization

This approach of vectorization comes close to the commonly used method for applying SOM on a set of documents. Each of the topic characteristics is converted into words and these in turn are counted.

3.1.1 Names and Occurrences

For names and occurrences their value and type are split up into words. Each word is a feature and counted as such. This means that multiple names or occurrences of the same type will cause a higher count for the features describing the words in these characteristics. There is only one exception to this rule, the name of the default name type is not counted as a feature for the topic because it adds no additional semantic value.

This thesis has not used any locator based values for occurrences and has not included a method of conversion to words for these values.

3.1.2 Associations

Associations contain the most semantic value in a topic map. They have a type and one or more roles to indicate direction. Each role also has a type and a player. To convert an association into a set of features all the values of the types and players are collected and broken into words. The role in which the current topic is a player is skipped. This is done to avoid its name being over-valued as a feature.

3.1.3 Reification

Reification of a topic's characteristic is done by a topic. Therefore, the words in the name of the reifying topic can be used as features too. In most situations the reifying topic has a name that describes the association and will therefore activate features that signal this specific reification. Through this the semantic value of the reification is persisted in the topic vector.

3.1.4 Scoping

The scope of a topic's characteristic is a set of topics. So, as with reification, the words in the names of these topics can be used as features. This means that the semantic value of the scope is placed at the topic, not at the characteristic. This is not a problem because all the characteristics are combined into a set of words. The significance of the scope is preserved because it will still increase values of their features. For example, if the scoping topic 'voice' is used on many associations for a topic then the vector for this topic will reflect this with the value of the 'voice' feature being higher.

3.2 Ontology based vectorization

The values of the names, occurrences and association players are not the only distinct properties of a topic. The actual number of names, occurrences and associations they have can also be seen as a unique footprint. This approach focuses on those ontological properties of a topic.

As with data based vectorization, there are two sets of rules to apply to the characteristics of a topic. However, there is a generalization. During calculation of the vectors a set of features is maintained for the entire document space. A feature within the ontologic vectorization is a quintuple consisting of an n-tuple of topics (F), a reification flag (R), a set of scoping topics (S), a document frequency (df) and a term frequency (tf), as shown in (3.1).

$$Feat = (F, R, S, df, tf) \quad \text{where} \quad |F| \in \{1, 3\} \quad \text{and} \quad R \in \{0, 1\} \tag{3.1}$$

To assure that the vector for a topic preserves any scope and reification semantic value the equality of features is determined based on the equality of F, R and S as shown in (3.2).

$$(Feat_x = Feat_y) \iff \begin{cases} (\forall x \in F_{Feat_x} : x \in F_{Feat_y}) & \wedge \\ (R_{Feat_x} = R_{Feat_y}) & \wedge \\ (\forall y \in S_{Feat_x} : y \in S_{Feat_y}) \end{cases} \tag{3.2}$$

3.2.1 Names and Occurrences

Both names and occurrences can be described as a pair containing a type and a value. Each one of these types is counted through the names and occurrences of a topic. A feature for a name-, or occurrence-type contains exactly one topic in F: the type of the name/occurrence, as shown in (3.3) and (3.4).

$$F_{name} = (Type_{name}) \tag{3.3}$$

$$F_{occurrence} = (Type_{occurrence}) \tag{3.4}$$

3.2.2 Associations

Associations introduce much more complexity when it comes to preserving the semantic value in the ontologic vectorization. The vectorization has to consider the association type and the role types. On top of that, the direction the association is viewed from adds even more semantic value.

To ensure feature uniqueness regarding all these conditions, an association's feature's F contains the association type, the current topic's role type and the set of other role types, as shown in (3.5). This approach guarantees the feature is unique for each number of roles used in combination with the association type and that the direction is also unique.

$$F_{association} = (Type_{association}, Type_{thisRole}, \{Type^1_{otherRole}, \ldots, Type^n_{otherRole}\}) \tag{3.5}$$

Note that the third element of F here is a set.

Consider the association in Figure 3.1 as an example. Assume that both $T1$ and $T2$ are focus topics in need of vectorization. When viewed from in the context of $T1$ the F value of the feature for this association would be:

$$F_{T1} = (at, rt1, \{rt2\}) \tag{3.6}$$

And in the context of $T2$:

$$F_{T2} = (at, rt2, \{rt1\}) \tag{3.7}$$

These F values are not equal. Therefore, directional context has provided two features for the same association.

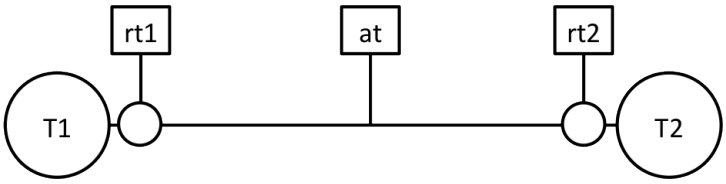


Figure 3.1: An example association with two players and distinct role types

3.2.3 Reification

If a characteristic is reified by a topic the feature will reflect this by setting the R flag to 1. If no reification is present this flag will be set to 0. Due to the equality rule specified above features of equal type and scope will be unique based on their reification value and introduce a separate vector value.

3.2.4 Scoping

As mentioned in the introduction, scoping adds a lot of semantic value to a characteristic in the form of context. For the data based vectorization this context reflects the topic. Yet, for the ontologic vectorization this context needs to reflect the characteristic itself. Therefore, it is included in the feature definition as proposed by (3.1).

The result of this inclusion is that two associations that are equal except for their scope will be mapped on two distinct features.

Chapter 4

Experiments

Applying the vectorization methods proposed in Chapter 3 to the test set as described in Section 2.3 lead to 20 sets of vectors. These vectors were used to train the SOM and GHSOM networks by using the Java SOMToolbox.

This chapter first introduces the Java SOMToolbox. The following sections explain which SOM techniques were used and how the results are visualized. Finally, the results of the experiments for the static SOM and the GHSOM techniques are presented and explained.

4.1 The Java SOMToolbox

The Java SOMToolbox was developed at Institute of Software Technology and Interactive System the Vienna University of Technology. The toolbox includes several implementations of SOM techniques, visualizations, comparisons techniques and quality measurements. [6]

The toolbox has clear definitions for input, output and supporting files for which the vectorization methods were adapted.

4.2 Tested SOM techniques

The 20 data sets were used as a training set for a static SOM and a GHSOM resulting in 40 trained SOM networks. The parameters used for the training were kept on the default values as specified by the SOMToolbox. Differences between the techniques in these parameters were kept to the bare necessities to give both training techniques equal chances of successful clustering.

4.3 Visualization

The Thematic Class Map visualization [13] was used to visualize and analyse the SOM maps. This visualization presents the clustering of the SOM regarding

the known classification depicted in eight colors. This visualization gives a clear view over the performance of the SOM because clusters of the same class can easily be recognized.

The thematic class map visualization is based on a voronoi diagram [1] and the SOMToolbox supports showing the voronoi lines. However, for this thesis the voronoi lines are not depicted in the visualizations of the tests because the main interest lies with the clustering.

4.3.1 Error visualization

To visualize the mean quantitative error (MQE) of the test results a simple white-to-black gradient is used. Black SOM units have the highest MQE value; white units the lowest.

4.4 Static SOM results

Figure 4.1 and Figure 4.2 show the resulting static SOMs of all the test topic maps. From these results it is clear that both vectorization approaches have no trouble dealing with the main topic types in the test set when applied on a static SOM. The ontology based vectorization leads to slightly more complete clusters regarding the main topic types. This does not indicate failure on the data based vectorization part because the differences between types is more discrete in the ontology compared to words. The data based vectorization might have discovered previously unknown relations between the topics in these cases.

The problem with the less represented topics is just that: the SOM has not seen enough instances of those types during training to be able to provide a good generalization. This means the complete IMDb dataset is unlikely to result in a SOM in which every type is clearly clustered. The topic types that represent optional values — like genre, language, country, gender, etc. — play a big role with clustering of the larger types, such as characters, movies and episodes, but due to their low instance count are unlikely to cluster.

Table 4.1 displays the class legend for the result maps.

Actor	
Character	
Episode	
Movie	
Note	
Series	
TV	
Video	

Table 4.1: Class legend for result maps

Test set	0	1	2	3	4	5	6	7	8	9
Depth data based	14	12	13	12	14	13	13	13	13	14
Depth ontology based	4	3	4	4	3	4	4	4	4	4
Breadth data based	14	18	13	9	11	11	12	13	9	12
Breadth ontology based	8	8	7	9	8	7	8	7	8	7

Table 4.2: Test sets GHSOM dimensions

4.5 GHSOM results

As with the static SOM the top layer of the GHSOM clearly clusters the large topic types. The GHSOM does indicate a smaller static SOM network might be used to cluster the large topic types.

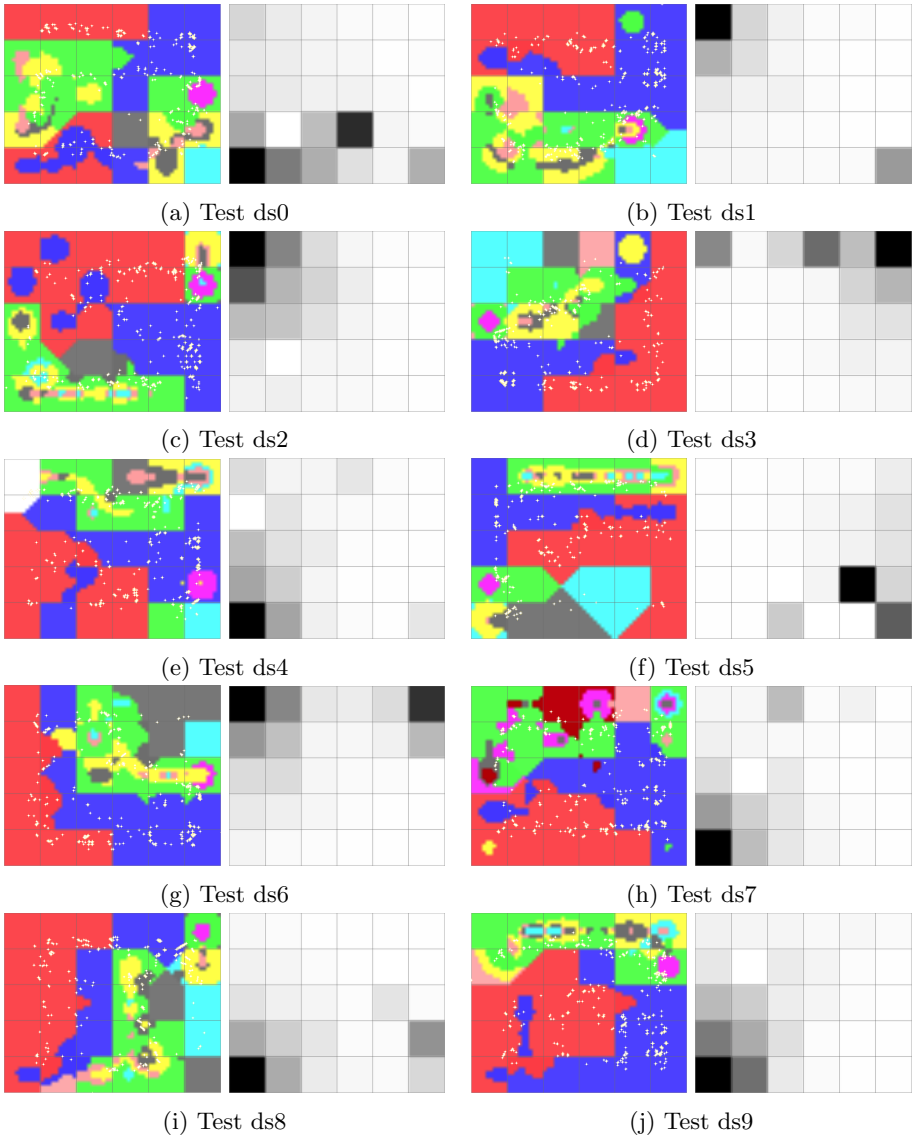
Due to the nature of the GHSOM a closer inspection of the sublayers is needed in order to determine the effectiveness of the hierarchical aspect of the network. Figure 4.5 displays the first two levels of the GHSOM of test topic map 0 trained on the data based vectorization. In some cases, like with submap $(2, 0)$[1] the hierarchy is very clear and stops going deeper.

In case of the $(1, 1)$ unit the GHSOM has expanded the unit with only characters for two more levels. This indicates that the GHSOM detected previously unknown differences between the characters. Whether these differences are worthy of further research depends on the intended level of abstraction.

Finally, the $(1, 0)$ is an example of the last case. This unit is expanded to a SOM that still represents a lot of different types. Most of the units in this SOM are further expanded into level 3 as (partially) shown in Figure 4.6. One of these expansions continues into the 14^{th} level as shown in Figure 4.7. This situation illustrates the power of the GHSOM's ability to find a hierarchy in a single topic type.

Comparing the data based vectorization with the ontology based vectorization when a GHSOM is used shows that the ontology based vectorization has a higher level of abstraction. Overall, the ontology based vectorization results have a lower depth and breadth then the data based vectorization results as shown in Table 4.2. The reason for this is that the ontology based vectorization results converge quicker due to fewer features.

[1] Sub maps are named (x, y) from the top left which is $(0, 0)$



(a) Test ds0 (b) Test ds1

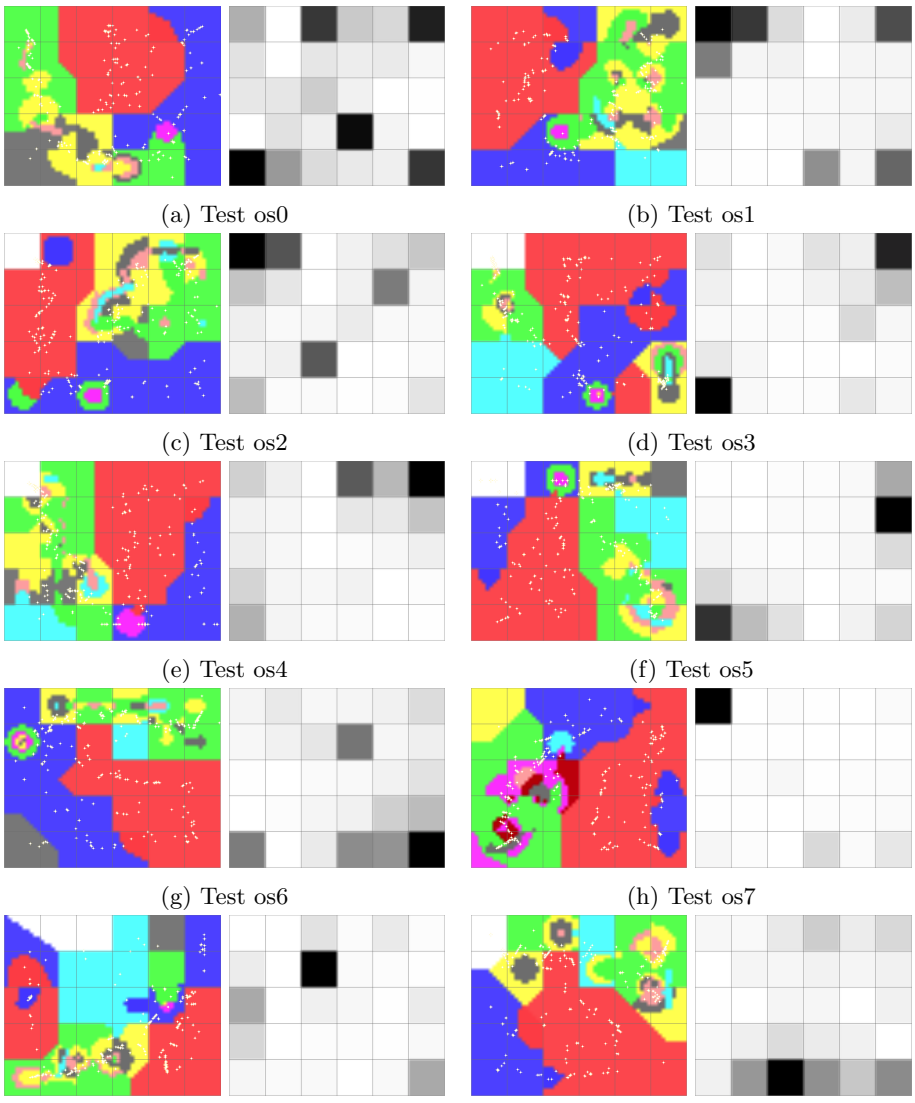
(c) Test ds2 (d) Test ds3

(e) Test ds4 (f) Test ds5

(g) Test ds6 (h) Test ds7

(i) Test ds8 (j) Test ds9

Figure 4.1: Data based vectorization applied to a static SOM network



(a) Test os0 (b) Test os1

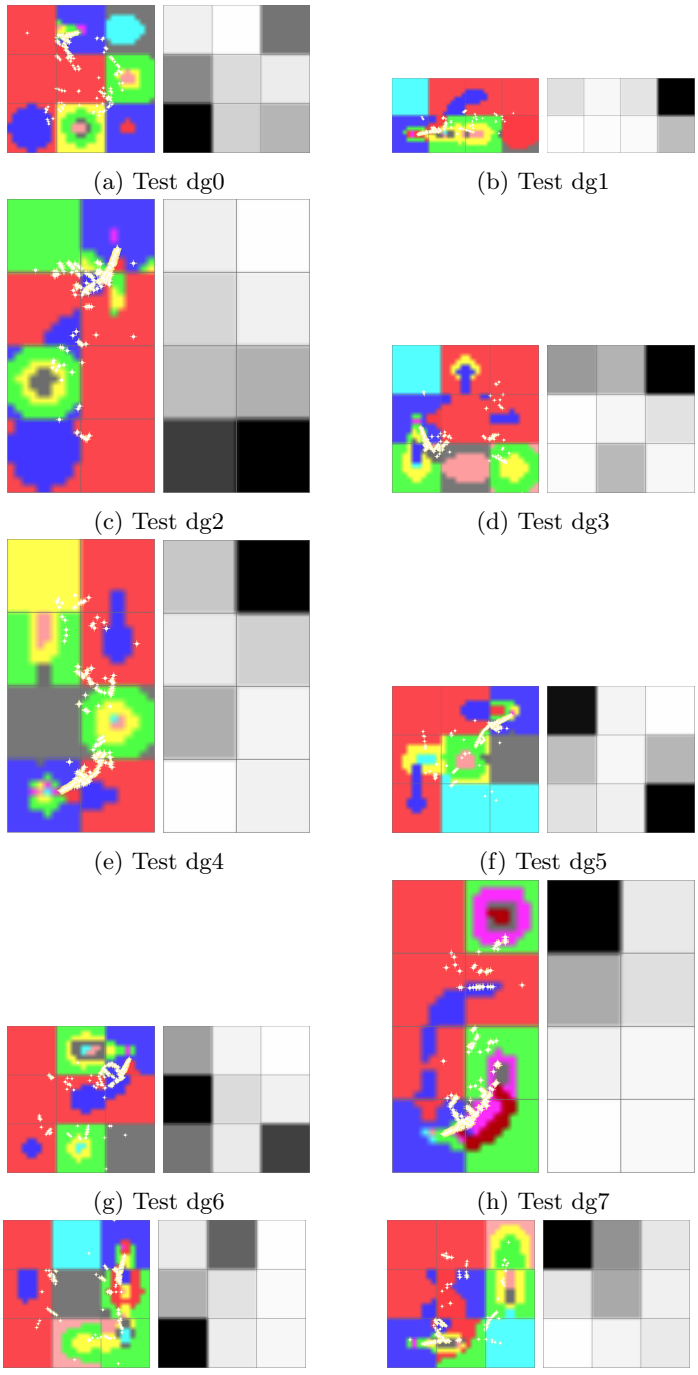
(c) Test os2 (d) Test os3

(e) Test os4 (f) Test os5

(g) Test os6 (h) Test os7

(i) Test os8 (j) Test os9

Figure 4.2: Ontologic based vectorization applied to a static SOM network



(a) Test dg0 (b) Test dg1

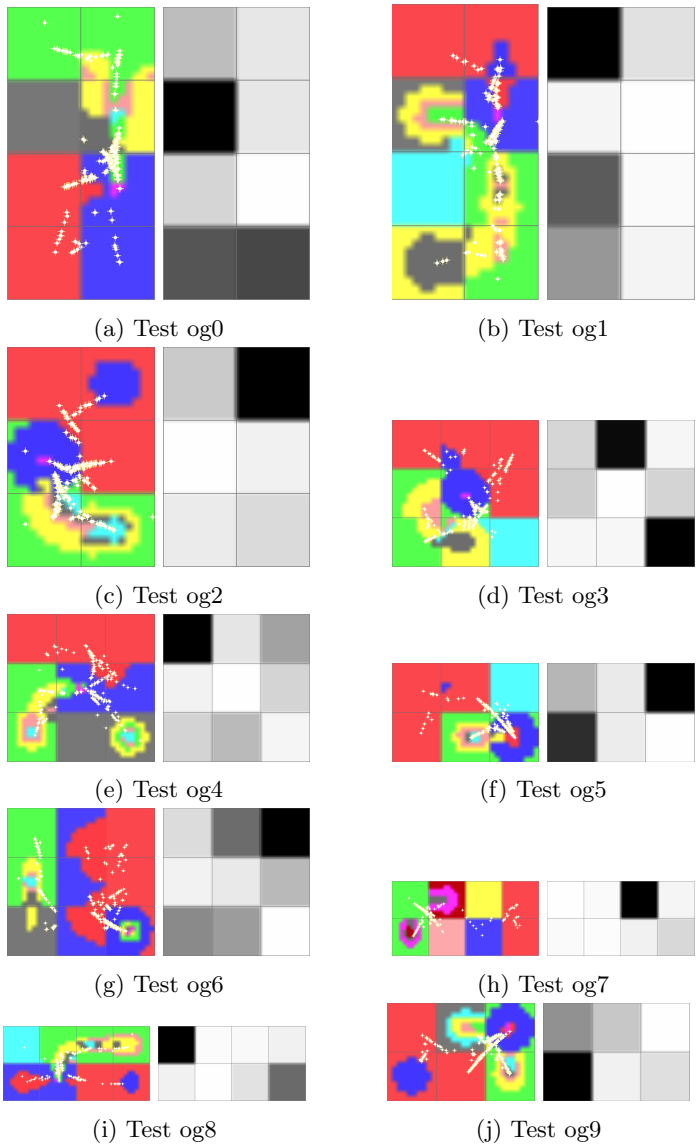
(c) Test dg2 (d) Test dg3

(e) Test dg4 (f) Test dg5

(g) Test dg6 (h) Test dg7

(i) Test dg8 (j) Test dg9

Figure 4.3: Data based vectorization applied to a GHSOM network (top layer)



(a) Test og0 (b) Test og1

(c) Test og2 (d) Test og3

(e) Test og4 (f) Test og5

(g) Test og6 (h) Test og7

(i) Test og8 (j) Test og9

Figure 4.4: Ontologic based vectorization applied to a GHSOM network (top layer)

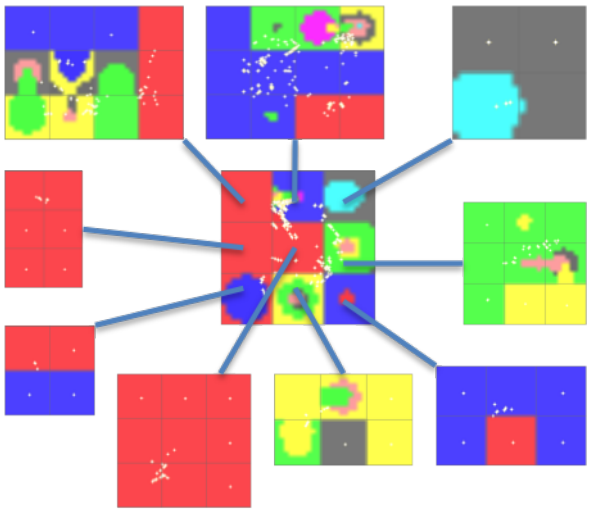


Figure 4.5: First two levels of the GHSOM of data based vectorization for test topic map 0

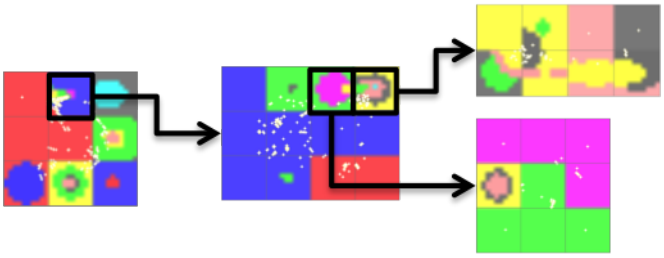


Figure 4.6: Partial path through the GHSOM layers for the $(1, 0)$ unit of layer 1 in data based vectorization test topic map 0

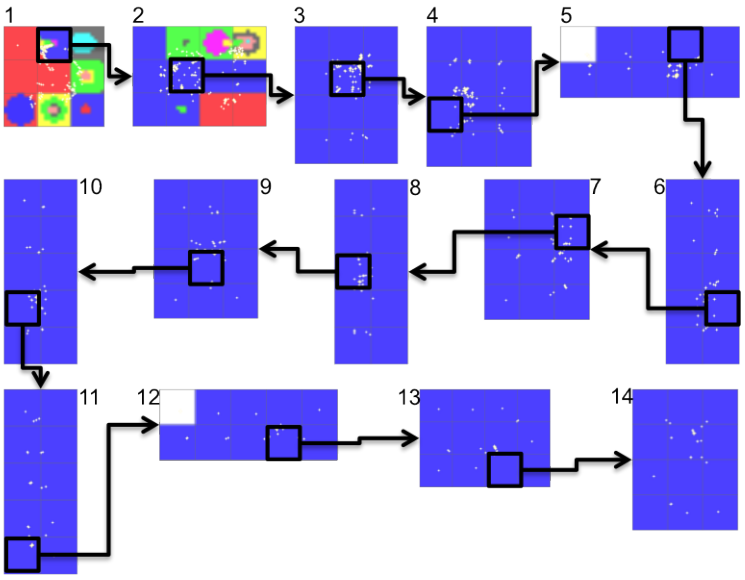


Figure 4.7: Partial path through 14 GHSOM layers for the $(1, 0)$ unit of layer 1 in data based vectorization test topic map 0

Chapter 5

Conclusion

This chapter gives the conclusion for this thesis by first answering the research questions and finally answering the problem statement.

5.1 Research questions answers

The first two research questions proposed in Section 1.4 focus on the vectorization process.

Research question 1 Can a vectorization method be defined that uses the textual values stored in a topic map?

This method of vectorization was proposed in Chapter 3 as the data based vectorization method. As discussed in Chapter 4, the SOM networks trained with the data based vectorization preformed as expected. The GHSOM provided a deeper hierarchy in the topics as it detected relationships between the words. Therefore, we can answer this research question with a clear 'yes'. Furthermore, the topic map has enough textual data for a SOM to work with.

Research question 2 Can a vectorization method be defined that uses the ontological properties of a topic?

This method of vectorization was also proposed in Chapter 3, as the ontology based vectorization. Although a completely different approach compared to the data based vectorization, it performed equally well. The results in Chapter 4 show that the GHSOM technique does not provided a very deep hierarchy when the ontology based vectorization is used. However, this does not change the answer of the research question. Yes the ontology based vectorization can be used as a vectorization of topic map based data.

Research question 3 Can the static SOM technique be applied to the two proposed vectorization methods?

Research question 4 Can the GHSOM technique be applied to the two proposed vectorization methods?

The answers to these two questions are equal. Both techniques can be applied to the proposed vectorization methods. The difference in the level of detail between the technique and vectorization method combinations indicate that an optimal combination exists for a arbitrary problem domain.

5.2 Problem statement answer

The problem statement proposed in Section 1.4 was:

Problem statement How can existing SOM techniques be applied to semantic data represented in a topic map?

This problem can now be answered from the research question answers. The static SOM and GHSOM techniques can be applied on topic maps data by using the proposed vectorization methods. Both vectorization methods result in a clustering in both tested SOM techniques. The data based vectorization will provide a deeper hierarchy when used in combination with a GHSOM. The ontology based vectorization can be used if this is not wanted or textual values are not sufficiently at hand.

Chapter 6

Future work

This thesis focused on a specific problem regarding the combination of SOM networks and semantic data. There are three possible additional research directions that can be done to improve or build on the results of this thesis.

6.1 Complex semantic dataset

The IMDb corpus, although large, focuses mainly on titles. Therefore the resulting topic map is large, but not very complex. The network of topics and associations is mainly build on the actor-title associations and the supporting types. The proposed methods could also be tested on a more complex topic map.

6.2 Parameter estimation

As specified in Section 4.2 the training of the SOM networks was done with default parameters. The clustering results might be improved by choosing different parameter values.

6.3 Automatic classification

A possible future research would be the application of the proposed vectorization methods in a (semi)automatic classification process. A SOM could be trained and analysed for a sample from the complete dataset. This trained SOM could then be used to quickly classify the remainder of the dataset.

Bibliography

[1] F. Aurenhammer and R. Klein. Voronoi diagrams. In *Handbook of Computational Geometry*, pages 201–290. Elsevier Science Publishers B.V. North-Holland.

[2] M. Dittenbach, D. Merkl, and A. Rauber. The growing hierarchical self-organizing map. In *Proceedings of the IEEE-INNS-ENNS International Joint Conference on Neural Networks. IJCNN 2000. Neural Computing: New Challenges and Perspectives for the New Millennium*, pages 15–19. IEEE Comput. Soc, 2000. ISBN 0-7695-0619-4. URL `http://dx.doi.org/10.1109/IJCNN.2000.859366`.

[3] L. Garshol. A theory of scope. In *Scaling Topic Maps*, volume 4999 of *Lecture Notes in Computer Science*, pages 74–85, 2008.

[4] ISO/IEC 13250:2003. *Information Technology - Document Description and Processing Languages - Topic Maps*. International Organization for Standardization, Geneva, Switzerland.

[5] ISO/IEC IS 13250-2:2006. *Information Technology - Document Description and Processing Languages - Topic Maps - Data Model*. International Organization for Standardization, Geneva, Switzerland.

[6] Institute of Software Technology and Interactive Systems, Vienna University of Technology. The Java SOMToolbox, 2011. URL `http://www.ifs.tuwien.ac.at/dm/somtoolbox`.

[7] International Organization for Standardization. Information processing - text and office systems - standard generalized markup language (SGML). Technical Report ISO 8879, International Organization for Standardization, Geneva, Switzerland, 1986.

[8] Internet Movie Database. About imdb. URL `http://www.imdb.com/pressroom/about/`.

[9] Internet Movie Database. Alternative interfaces. URL `http://www.imdb.com/interfaces/`.

[10] Internet Movie Database. Company timeline. URL http://www.imdb.com/pressroom/company_timeline/.

[11] S. Kaskiy, J. Kangasz, and T. Kohonen. Bibliography of self-organizing map (som) papers: 1981-1997. *Neural Computing Surveys*, 1:102–350, 1998.

[12] T. Kohonen. The self-organizing map. *Neurocomputing*, 21(1-3):1–6, 1998. ISSN 09252312.

[13] R. Mayer, T. Abdel Aziz, and A. Rauber. Visualising class distribution on self-organising maps. In *Proceedings of the International Conference on Artificial Neural Networks (ICANN'07)*, volume 4669 of *LNCS*, pages 359–368. Springer.

[14] J. Sowa. *Knowledge Representation: Logical, Philosophical and Computational Foundations*. Brooks / Cole, Pacific Grove, 2000.

[15] Wikipedia. Internet movie database, August 2012. URL http://en.wikipedia.org/wiki/Internet_Movie_Database.

[16] World Wide Web Consortium. Extensible markup language (xml), 2006. URL http://www.w3.org/XML/.

[17] World Wide Web Consortium. Resource description framework (rdf), 2006. URL http://www.w3.org/RDF/.