(Missing) Concept Discovery in Heterogeneous Information Networks

Tobias Kötter and **Michael R. Berthold**
Nycomed Chair for Bioinformatics and Information Mining
University of Konstanz, Germany
Tobias.Koetter@uni-konstanz.de

Abstract

This paper proposes a new approach to extract existing (or detect missing) concepts from a loosely integrated collection of information units by means of concept graph detection. Once the concepts have been extracted they can be used in order to create a higher level representation of the data. Concept graphs further allow the discovery of missing concepts which might lead to new insights by connecting seemingly unrelated information units.

Introduction

The amount of data researchers have access to increases at a breath taking pace. The available data stems from heterogeneous sources from diverse domains with varying semantics and of various quality. It is a big challenge to integrate and reason from such an amount of data. However by integrating data from diverse domains one might discover relations that span across multiple domains leading to new insights and thus a better understanding of complex systems. In this paper we use a network-based approach to integrate data from diverse domains of varying quality. The network consists of vertices that represent information units such as objects, ideas or emotions, whereas edges represent the relations between these information units.

Once the data has been merged into a unifying model it needs to be analyzed. In this paper we propose concept graphs as an approach to extract semantical information from loosely integrated information fragments. Concept graphs allow for the detection of existing concepts which can be used to create an abstraction of the underlying data. By providing a higher level view on the data the user might get a better insight into the integrated data and discover new relations across diverse domains that have been hidden in the noise of the integrated data.

Concept graphs also allow for the detection of domain bridging concepts (Kötter, Thiel, and Berthold 2010) that connect information units from various domains. Domain bridging concepts might support creative thinking by connecting seemingly unrelated information units from diverse domains.

Another advantage of concept graphs is that they enable the detection of information units that share common properties but to which no concept has been assigned yet. This might lead to the discovery of concepts that are missing in the data or to the detection of new concepts.

The rest of the paper is organized as follows: in the next chapter we will briefly review Bisociative Information Networks, which we use for the integration of heterogeneous data sources from diverse domains. Subsequently we will introduce concept graphs and describe their detection. We will then discuss the discovery of concept graphs in a real world data set and show some example graphs. Finally we draw conclusions from our discussion and give an outlook on future work.

Bisociative Information Networks

Bisociative Information Networks (BisoNets) (Berthold et al. 2008) provide a framework for the integration of semantically meaningful information but also loosely coupled information fragments from heterogeneous data sources. The term *bisociation* (Koestler 1964) was coined by Arthur Koestler in 1964 to indicate the "...joining of unrelated, often conflicting information in a new way...".

BisoNets are based on a k-partite graph structure, whereby the most trivial partitioning would consist of two partitions ($k = 2$), with the first vertex set representing units of information and the second set representing the relations among information units. By representing relations as vertices BisoNets support the modeling of relationships among any number of members.

However the role of a vertex is not fixed in the data. Depending on the point of view a vertex can represent an information unit or a relation describing the connection between units of information. Members of a relation are connected by an edge with the vertex describing the relation they share. One example is the representation of documents and authors where documents as well as authors are represented as vertices. Depending on the point of view, a document might play the role of the relation describing authorship or might be a member in the relation of documents written by the same author.

The unified modeling of information units and relations as vertices has many advantages e.g. they both support assigning of attributes such as different labels. However these attributes do not carry any semantic information. Edges can be further marked as directed to explicit model relationships that are only valid in one direction. Vertices can also be assigned to partitions to distinguish between different domains such as biology, chemistry, etc.

Since relations are assigned a weight that describes the reliability of the connection, in contrast to ontologies, semantic networks or topic maps BisoNets support the integration of not only facts but also pieces of evidence. Thus units of information and their relations can be extracted from various information sources such as existing databases, ontologies or semantical networks. But also semistructured and noisy data such as literature or biological experiments can be integrated in order to provide a much richer and broader description of the information units. By applying different mining algorithms on the same information source diverse relations and units of information can be extracted, where each mining algorithm represents an alternative view that might highlight a different aspect of the same data.

BisoNets focus only on the information units and their relations alone without storing all the more detailed data underneath the pieces of information. However vertices do reference the detailed data they stem from. This allows BisoNets to integrate huge amounts of data and still be able to show the data from which a vertex originates.

Concept Graphs

Once all the data has been integrated, it has to be analyzed in order to find valuable information. We propose a new method to extract semantical information from the loosely integrated collection of information units by the means of concept graph detection.

A concept graph represents a *concept* which stands for a mental symbol. A concept consists of *information units*, which do not only refer to materialized objects but also to ideas, activities or events, and also their shared *aspects*, which represent the properties the information units share. In philosophy and psychology, information units are also known as the extension of a concept, which consists of the things to which the concept applies. Whereby the aspects are known as the intension of a concept, consisting of the idea or the properties of the concept. An example would be a concept representing birds with specific birds such as eagles or sparrows as information units, which in turn are related to their common aspects such as feather, wing, and beak.

In addition to the information units and their shared aspects, a concept graph might also contain the symbolic representation of the concept itself. This symbolic representation can be used to generate an abstract view on the data since it represents all members of the corresponding concept graph.

An example of a concept graph that represents the concept of *flightless birds* is depicted in Figure 1. It consists of the two information units *Ostrich* and *Weka* and their shared aspects *wing* and *feather*. The graph also contains the symbolic representation of the flightless bird concept, which can be used as an abstract representation of this particular concept graph.

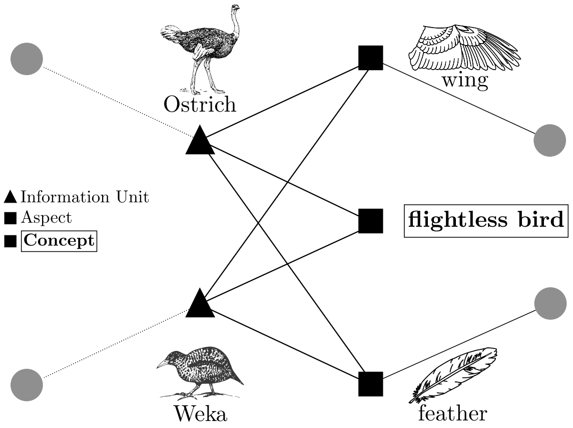


Figure 1: Example concept graph

Preliminaries

As mentioned above a concept graph contains information units which are similar in that they share some aspects. In BisoNets the aspects of an information unit are represented by its direct neighbors. The more neighbors two information units share the more similar they are. This leads to the representation of a concept graph as a dense subgraph in a BisoNet, consisting of two disjoint and fully connected vertex sets. Here the first vertex set represents the information units and the second vertex set the aspects that are shared by all information units of the concept graph. Thus a perfect concept graph would form a complete bipartite graph as depicted in Figure 1 with the information units as the first partition and the aspects with the concept as the second partition. An imperfect concept graph also contains relations among the vertices within a partition and thus does not form a perfect bipartite (sub) graph. However, such imprecise concept graphs are of prime interest, of course.

Once a dense subgraph has been detected it needs to be analyzed in order to distinguish between the information unit set and the aspect set. We have developed heuristics to detect the different set types for directed and undirected networks. Both heuristics are based on the assumption that information units are described by their neighbors in the network. The heuristics for the directed network are also based on the assumption that information units point to their aspects. Hence in a directed network a relation consists of an information unit as source and an aspect as target vertex.

The heuristics to identify the different vertex types are based on the following definitions:

Let $B(V, E)$ be the un/directed BisoNet that contains all information with V representing the vertices and $E \subseteq V \times V$ representing the edges. $C(V_A, V_I, E') \subseteq B$ defines the concept graph C in the BisoNet B. $V_A \subseteq V$ represents the aspect set and $V_I \subseteq V$ the information unit set of the concept graph C in which $V_A \cap V_I = \emptyset$. $E' \subseteq E$ is the set of edges that fully connects the vertex sets of the concept graph so that $V_A \times V_I \subseteq E'$.

Let

$$N(v) = \{u \in V : \{v, u\} \in E\}$$

be the neighbors of the vertex $v \in V$ in the BisoNet B.

Whereby

$$N^{+}(v) = \{u \in V : (v,u) \in E\}$$

denotes its target neighbors and

$$N^{-}(v) = \{u \in V : (u,v) \in E\}$$

its source neighbors.

The neighbors within a concept graph C for a vertex $v \in V_A \cup V_I$ are denoted by

$$N_C(v) = \{u \in V_A \cup V_I : \{v,u\} \in E'\}.$$

While

$$N_C^{+}(v) = \{u \in V_A \cup V_I : (v,u) \in E'\}$$

denotes its target neighbors and

$$N_C^{-}(v) = \{u \in V_A \cup V_I : (u,v) \in E'\}$$

its source neighbors.

Information unit set The information units form the first of the two disjoint vertex sets of the concept graph. The heuristic that denotes the probability of a vertex set to be the information unit set is denoted by the function $\mathrm{i}(V') \to [0,1], V' \subseteq V$.

In an undirected network $\mathrm{i}(V')$ is defined as the product of the ratios of neighbors inside and outside the concept graph for each vertex in V'

$$\mathrm{i}(V') = \prod_{v \in V'} \frac{|N_C(v)|}{|N(v)|}.$$

In a directed network the heuristic is defined as the product of the ratios of target neighbors within and outside of the concept graph for each vertex in V'

$$\mathrm{i}(V') = \prod_{v \in V'} \frac{|N_C^{+}(v)|}{|N^{+}(v)|}.$$

The information unit set $V_I \subseteq V$ is the vertex set of the concept graph that maximizes the function $\mathrm{i}(V')$.

Aspect set The aspect set is the second vertex set of the concept graph that describes the information units of the concept graph. Each aspect on its own might be related to other vertices as well but the set of aspects is only shared by the information units of the concept graph. The members of the aspect set might differ highly in the number of relations to vertices outside of the concept graph depending on their level of detail. More abstract aspects such as animals are likely to share more neighbors outside of the concept graph than more detailed aspects such as bird.

The heuristic that denotes the probability of a vertex set to belong to the aspect set is denoted by the function $\mathrm{a}(V') \to [0,1], V' \subseteq V$.

In an undirected network $\mathrm{a}(V')$ is defined as the product of the inverse ratios of neighbors inside and outside the concept graph for each vertex in V'

$$\mathrm{a}(V') = 1 - \prod_{v \in V'} \frac{|N_C(v)|}{|N(v)|} = 1 - \mathrm{i}(V').$$

In a directed network the heuristic is defined as the product of the ratios of the source neighbors inside and outside the concept graph for each vertex in V'

$$\mathrm{a}(V') = \prod_{v \in V'} \frac{|N_C^{-}(v)|}{|N^{-}(v)|}.$$

The aspect set $V_A \subseteq V$ is the vertex set of the concept graph that maximizes the function $\mathrm{a}(V')$.

Concepts The concept is a member of the aspect set V_A. A concept differs from the other members of the aspect set in that it should only be related to the information units within the concept graph. Hence a perfect concept has no relations to vertices outside of the concept graph and can thus be used to represent the concept graph.

The heuristic that denotes the probability of a vertex to be the concept that can represent a concept graph C is denoted by the function $\mathrm{c}(v) \to [0,1], v \in V_A$ whereby 1 denotes a perfect concept.

For an undirected network the heuristic is defined as the ratio of the neighbors inside and outside the concept graph

$$\mathrm{c}(v) = \frac{|N_C(v)|}{|N(v)|}.$$

In a directed network the heuristic considers the ratio of the source neighbors inside and outside the concept graph

$$\mathrm{c}(v) = \frac{|N_C^{-}(v)|}{|N^{-}(v)|}.$$

The concept that can represent the concept graph is the vertex v from the aspect set V_A with the highest value for $\mathrm{c}(v)$.

Depending on a user-given threshold we are able to detect a concept graph without a concept. The concept graph lacks a concept if the concept value $\mathrm{c}(v)$ of all vertices of its aspect set is below the given threshold. This might be an indication of an unknown relation among information units that has not been discovered yet and to which no concept has been assigned.

Detection

In this paper we use a frequent item set mining algorithm (Agrawal and Srikant 1994) to detect concept graphs in BisoNets. By using frequent item set algorithms we are able to detect concept graphs of different sizes and specificity.

Frequent item set mining has been developed for market basket analysis in order to find sets of products that are frequently bought together. It operates on a transaction database that consists of a transaction identifier and the products that have been bought together in the transaction. Represented as a graph, the overlapping transactions form a complete bipartite graph, which is the basis of our concept graphs.

In order to apply frequent item set mining algorithms to find concept graphs in BisoNets we have to convert the network representation into a transaction database. Therefore, for each vertex in the BisoNet, we create an entry in the transaction database with the vertex as the identifier and its direct neighbors as the products. Once the database has been created we can apply frequent item set mining algorithms to detect vertices that share some neighbors.

Frequent item set mining algorithms allow the selection of a minimum support that defines the minimum number of transactions containing a given item set in order to make it frequent. They also allow a minimum size to be set for the item set itself in order to discard all item sets that contain fewer items than the given threshold. By setting these two thresholds we are able to define the minimum size of the concept graph.

Since we want to find concept graphs of different specificity we need an additional threshold that takes the general overlap of the transactions into account. To achieve this we used an adaption of the Eclat (Zaki et al. 1997) algorithm called Jaccard Item Set Mining (JIM) (Segond and Borgelt 2011). JIM uses the Jaccard index (Jaccard 1901) as an additional threshold for pruning the frequent item sets. For two arbitrary sets A and B the Jaccard index is defined as

$$\mathrm{j}(A, B) = \frac{|A \cap B|}{|A \cup B|}.$$

Obviously, $\mathrm{j}(A, B)$ is 1 if the sets coincide (i.e. $A = B$) and 0 if they are disjoint (i.e. $A \cap B = \emptyset$).

By setting the threshold for the JIM algorithm between 0 and 1 we are able to detect concept graphs of different specificity. By setting the threshold to 1 only those vertices that share all of their neighbors are retained by the algorithm. This results in the detection of more specific concept graphs which contain either information units or aspects that exclusively belong to the detected concept graph. Relaxing the threshold by setting a smaller value results in the detection of more general concept graphs where the information units share some but not all of their aspects. Varying thresholds might lead to the detection of overlapping concept graphs. This can be used to create a hierarchy among the concepts.

Application

The 2008/09 Wikipedia Selection for schools[1] (Schools Wikipedia) is a free, hand-checked, non-commercial selection of the English Wikipedia[2] funded by SOS Children's Villages. It has been created with the intention to build a child safe encyclopedia. It has about 5500 articles and is about the size of a twenty volume encyclopedia (34,000 images and 20 million words). The encyclopedia contains 154 subjects which are grouped into 16 main subjects such as countries, religion and science. The network has been created from the Schools Wikipedia version created in October 2008. Each article is represented by a vertex and the subjects are represented by domains. Every article is assigned to one or more domains depending on the assigned subjects. Hyperlinks are represented by directed links with the article that contains the hyperlink as source and the referenced article as the target vertex.

This example data set and the representation as a hyperlink graph has been chosen since it can be validated manually by reading the Schools Wikipedia articles and inspecting their hyperlinks.

Results

This section illustrates concept graphs discovered in the Schools Wikipedia dataset using the JIM algorithm. The concept graphs consist of the discovered item sets that form the first vertex set and the corresponding root vertices of the transaction that build the second vertex set. Once we have discovered both vertex sets and determined their types we can display them as a graph.

The following graphs display the information units with triangular vertices. Both aspects and the concept are represented by a squared vertex whereas the concept has a box around its label.

Figure 2 depicts two different bird categories which were extracted from the animal section of the Schools Wikipedia dataset. Both graphs depict the aspects and the concept in their center and the information units in the surrounding circle.

The first concept graph (Figure 2a) represents the group of waders. Waders are long-legged wading birds such as herons, flamingos and plovers. The concept graph also contains terns and gulls even though they are only distantly related to waders. However Schools Wikipedia states that studies in 2004 showed that some of the gene sequences of terns showed a close relationship between terns and the Thinocori some species of aberrant waders. Reptiles are included in the graph since most of the larger waders eat reptiles.

The second concept graph (Figure 2b) represents the bird of prey group. Birds of prey or raptors hunt for food on the wing. The graph includes all the different sub families such as eagle, hawk, kite, osprey and falcon. It also includes some of the birds' prey such as chicken or crows. The common cuckoo is not a bird of prey but is included in the concept graph since it looks like a small bird of prey in flight as stated in its article in Schools Wikipedia.

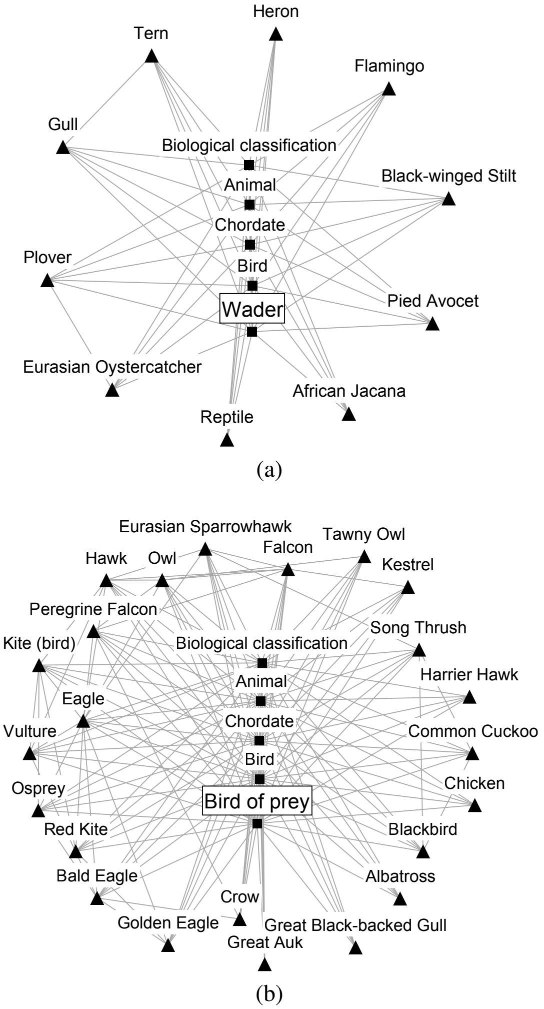
The animal examples benefit from the structure of the Schools Wikipedia pages of the animal section. They all contain an information box with the Kingdom, Phylum etc. of the animal. However this demonstrates that our method is able to discover ontologies if they are available in the integrated data. Furthermore the examples demonstrate the capability of the method to detect specific categories such as waders or birds of prey even though they are not part of the ontology structure in Schools Wikipedia.

In contrast to the animal graphs do the next concept graphs contain more aspects than information units. Therefore the layout of the vertices has changed. The information units are depicted in the center whereas the aspects and the concepts form the outer circle.

Figure 3 stems from the math section of the Schools Wikipedia data set and demonstrates the ability to detect specific concepts only based on the shared properties without an integrated ontology.

[1] http://schools-wikipedia.org/

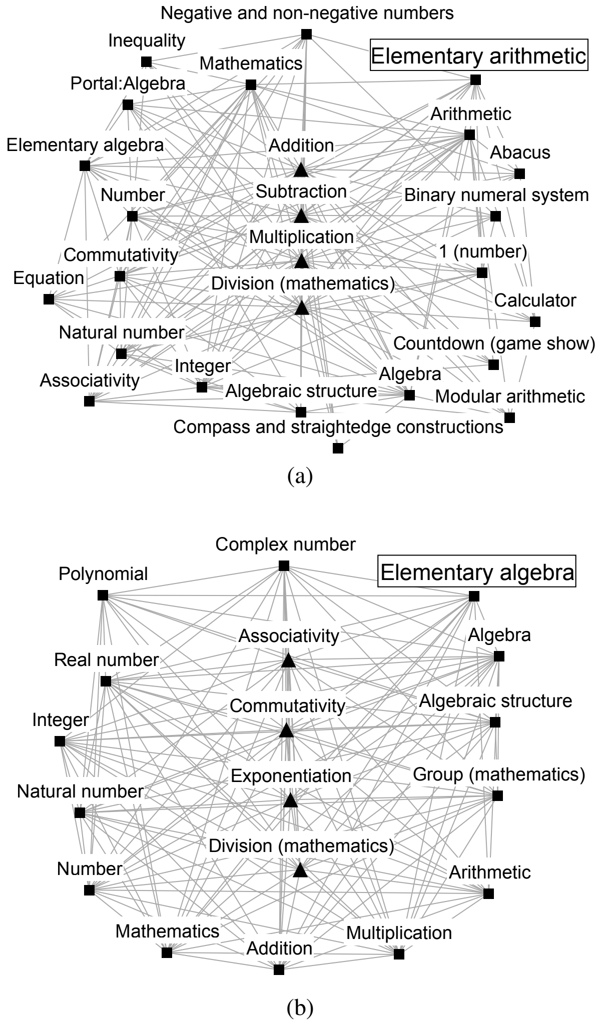
[2] http://en.wikipedia.org



(a)

(b)

Figure 2: Concept graphs from the animal section



(a)

(b)

Figure 3: Concept graphs from the mathematical section

The first concept graph (Figure 3a) represents the concept of elementary arithmetic, and grouping the main operations of elementary arithmetic such as addition, subtraction, multiplication and division. It also contains the vertex for abacus since some of them, such as the Chinese suanpans, can be used to perform all of the mentioned elementary arithmetic operations. The graph also contains the vertex for the elementary algebra concept that extends elementary arithmetic by introducing symbols in addition to numbers. This is described in the following paragraph.

The second concept graph (Figure 3b) groups some of the main laws of elementary algebra such as commutativity and associativity. Distributive law and symbol are not part of the concept graph since they are not explicitly explained in Schools Wikipedia and therefore not linked in the article. This is a limitation on the used data but not on the method itself. This is why we want to incorporate more information about each article in the next version of the Schools Wikipedia, such as information from the full text of the articles using text mining methods.

Both math examples contain some common vertices that belong to more general concepts such as mathematics, arithmetic and algebra, which could be used to generate a hierarchy of the mathematical section of the Schools Wikipedia data set.

Figure 4 depicts two concept graphs from the physics domain of the Schools Wikipedia data set and demonstrates the detection of domain crossing concepts. The graphs do not contain previously unknown relations but cross several domains such as the domain for physics, astronomy, history, chemistry and people. The examples benefit to a certain extend, such as the animal examples from standardized information boxes in Schools Wikipedia.

The first concept graph (Figure 4a) represents the concept graph for quantum field theory. It groups information units from the astronomy, physics and people domains with the domain for history.

The second concept graph (Figure 4b) refers to the wave-particle duality concept, which combines the domains for physics, astronomy and people with the chemistry domain.

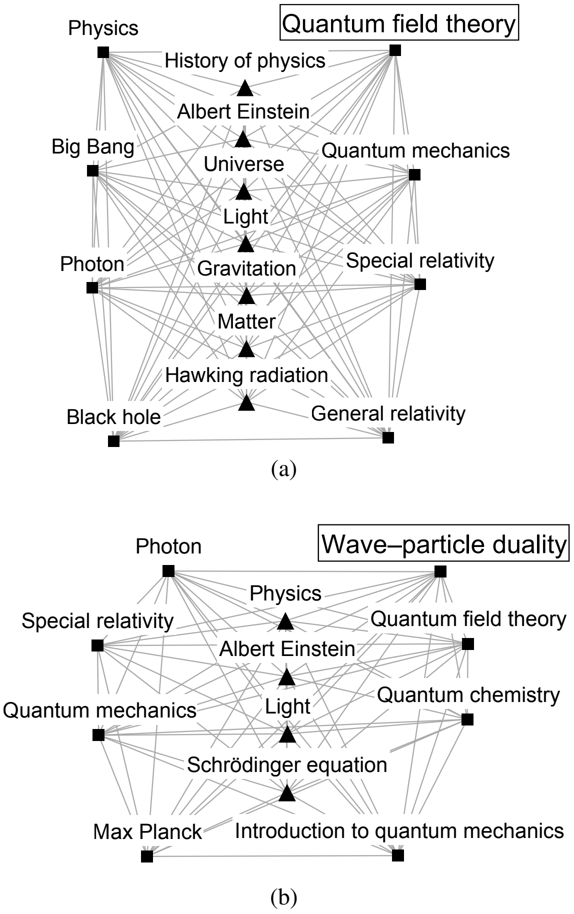


Figure 4: Domain bridging concept graphs from the physics section

Conclusion and Future work

In this paper we have discussed a new approach to detect existing or missing concepts from a loosely integrated collection of information fragments which leads to a deeper insight into the underlying data. We have discussed concept graphs as a way to discover conceptual information in BisoNets. Concept graphs allow for the abstraction of the data by detecting existing concepts leading to a better overview of the integrated data. They further support the detection of missing concepts by discovering information units that share certain aspects but which have no concept, which might be a hint for a previously unknown and potentially novel concept.

This approach can also be expanded to detect domain bridging concepts (Kötter, Thiel, and Berthold 2010) which might support creative thinking by connecting information units from diverse domains. Since BisoNets store the domain a vertex stems from, we can use this information to find concept graphs that contain information units from diverse domains.

In addition to the discovery of concept graphs we plan to identify overlapping concept graphs which can be used to create a hierarchy among the detected concepts using methods from formal concept analysis (Wille 1982). The hierarchy ranging from most specific to most general concepts can be created by detecting more specific concept graphs that are included in more general concept graphs. The different levels of concept graphs can be detected by varying the threshold of the discussed Jaccard Item Set Mining algorithm.

Acknowledgments

The work presented in this paper was supported by a European Commission grant under the 7th Framework Programme FP7-ICT-2007-C FET-Open, project no. BISON-211898.

References

Agrawal, R., and Srikant, R. 1994. Fast algorithms for mining association rules in large databases. In *Proceedings of the 20th International Conference on Very Large Data Bases*, VLDB '94, 487–499. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.

Berthold, M. R.; Dill, F.; Kötter, T.; and Thiel, K. 2008. Supporting creativity: Towards associative discovery of new insights. In *Proceedings of PAKDD 2008 (The Pacific-Asia Conference on Knowledge Discovery and Data Mining)*.

Jaccard, P. 1901. Étude comparative de la distribution florale dans une portion des alpes et des jura. *Bulletin de la Société Vaudoise des Sciences Naturells* 37:547–579.

Koestler, A. 1964. *The Act of Creation*. Macmillan.

Kötter, T.; Thiel, K.; and Berthold, M. R. 2010. Domain bridging associations support creativity. In *Proceedings of the International Conference on Computational Creativity, Lisbon*, 200–204.

Segond, M., and Borgelt, C. 2011. Item set mining based on cover similarity. In *Proceedings of PAKDD 2011 (The Pacific-Asia Conference on Knowledge Discovery and Data Mining)*.

Wille, R. 1982. Restructuring lattice theory: An approach based on hierarchies of concepts. In Rival, I., ed., *Ordered Sets*, 314–339. D. Reidel Publishing Company.

Zaki, M. J.; Parthasarathy, S.; Ogihara, M.; and Li, W. 1997. New algorithms for fast discovery of association rules. In *3rd Intl. Conf. on Knowledge Discovery and Data Mining*, 283–286. AAAI Press.