From Information Networks to Bisociative Information Networks

Tobias Kötter and Michael R. Berthold

Nycomed-Chair for Bioinformatics and Information Mining,
University of Konstanz, 78484 Konstanz, Germany
Tobias.Koetter@uni-Konstanz.de

Abstract. The integration of heterogeneous data from various domains without the need for prefiltering prepares the ground for bisociative knowledge discoveries where attempts are made to find unexpected relations across seemingly unrelated domains. Information networks, due to their flexible data structure, lend themselves perfectly to the integration of these heterogeneous data sources. This chapter provides an overview of different types of information networks and categorizes them by identifying several key properties of information units and relations which reflect the expressiveness and thus ability of an information network to model heterogeneous data from diverse domains. The chapter progresses by describing a new type of information network known as bisociative information networks. This kind of network combines the key properties of existing networks in order to provide the foundation for bisociative knowledge discoveries. Finally based on this data structure three different patterns are described that fulfill the requirements of a bisociation by connecting concepts from seemingly unrelated domains.

1 Introduction

Applications of bisociative creative information exploration derive their potential to produce creative discoveries, insight and solutions from exploring bisociations across large volumes of information originating from two or more domain theories. To facilitate such applications it is necessary to integrate these domain theories (or associated knowledge bases) in such a way that the integrated pool can be processed coherently. Integration of such data is a considerable challenge not only because of the data volumes, but also because of the semantic (ontologies of different domains) and syntactic (data and knowledge formats) heterogeneity involved.

An obvious approach to integrate these large volumes of information from various domains with varying quality is a flexible representation in terms of an information network. A number of different types of information networks have been proposed in the last few years [38] particularly in the area of biomedical domains. This area of research is known for its diverse information sources that need to be considered, for example, in the drug discovery process [12]. The integrated sources

M.R. Berthold (Ed.): Bisociative Knowledge Discovery, LNAI 7250, pp. 33–50, 2012.

range from experimental data, such as gene expression results, through to highly curated ontologies, such as the ontology of Medical Subject Headings[1].

Information networks are commonly composed of information units representing physical objects as well as immaterial objects such as ideas or events and relations representing semantic or solely correlational connections between information units. They are almost always based on a graph structure with vertices and edges, where vertices represent units of information, e.g. genes, proteins or diseases, and the relations between these units of information are usually represented by edges. In some information networks relations are represented by vertices as well, and therefore apply a bi-partite graph representation. This type of representation has the added advantage that relations between more than two information units can be easily supported. Furthermore an edge can be directed or undirected depending on the relationship it represents. Most networks also allow additional attributes or properties to be attached to vertices and edges, such as a vertex type, e.g. gene or protein, describing the nature of the information unit. Such information networks that connect multi-typed vertices are also known as heterogeneous information networks [28].

In order to integrate not only structured and well annotated repositories but also other types of information such as experimental data or results from text mining, some information networks support weighted edges. Therefore interactions in biological systems, which can be noisy and erroneous, are often modeled by Bayesian networks [22,24,31]. In these approaches the edge weight represents the probability of the existence of the connection. However, the edge weight of networks used by information retrieval techniques, such as knowledge or Hopfield networks [14], represents the relatedness of terms. Usually the weights in these approaches are computed only once. In contrast to these approaches, Belew enables each user of an adaptive information retrieval (AIR) model [6] to adapt the weights according to their relevance feedback. The disadvantage of this approach is that over time the network will be strongly biased by the opinions of the majority of the users. Another weighted-graph method constructs a weighted graph based on information extracted from available databases [49]. In doing so the edge weight represents the quality of the relation and is based on three factors: edge reliability, relevance and rarity. They assume that each edge type has a natural inverse, such as "coded by" and "is referred by". Similarly, there is one inverse edge for each edge, leading to an undirected graph with directed edge labels.

Once the data is represented in an information network this well-defined structure can be used to discover patterns of interest, extract network summarizations or abstractions and develop tools for the visual exploration of the underlying relations. A general analysis of the structure of complex networks stemming from real-world applications has been conducted by Albert and Barabasi [2]. They have discovered that these networks often share a number of common properties such as the small-world property, clustering coefficient or degree distribution. A survey on link mining has been conducted by Getoor and Diehl [27].

[1] `http://www.nlm.nih.gov/mesh/meshhome.html`

They classified the link mining task into three categories: object-related tasks, link-related tasks and graph-related tasks.

Network summarizations representing different levels of detail can be visualized to gain insight into the structure of the integrated data. A general introduction to network analysis can be found in [11]. An overview of existing graph clustering methods can be found in [48] and a review of graph visualization tools for biological networks can be found in [45]. The paper compares the functionality, limitation and specific strength of these tools.

Approaches from the semantic Web community include formalization of general semantic networks where the most popular variants have resulted in the RDF standard [40] and for formalism of topic maps [23]. Both techniques imply the construction of various formalizations in the form of different graph constructs. A highly complex example is the formalization of topic maps via *shifted hypergraphs* [3]. In this approach a hypergraph model for topic maps is defined in which the standard hypergraph is extended to a multi-level hypergraph via a shift function. RDF models were proposed in the form of different graph structures: graph [29], bipartite graph [30] and hypergraph [42]. Standard graphs allow the modeling of relations between two nodes, whereas bipartite graphs and hypergraphs permit the integration of relations among any number of members.

In order to visually analyze large networks with several million vertices and many more edges, visualization has to focus on a sub-graph or at least summarize the network to match the user's interest or provide some kind of overview of existing concepts. Various visualization and graph summarization techniques have been developed to address this problem. Examples can be seen in the generalized fisheye views [25], the interactive navigation through different levels of abstraction [1], the extraction of sub-graphs that contain most of the relevant information by querying [21] or by spreading-activation [18]. Other approaches summarize the graph by clustering or pruning it based on the topology [57] or additional information such as a given ontology [50].

The next section describes different types of information networks and characterizes them based on the features they support, which are relevant to the integration of heterogeneous data types. We subsequently introduce bisociative information networks, which have been tailored to support the integration of heterogeneous data sources. Before we move on to the conclusion, we discuss patterns of bisociation in this type of network that support creative thinking by connecting seemingly unrelated domains.

2 Different Categories of Information Network

In order to differentiate among information networks, distinctions can be made between different properties of information units and relations. These properties are, of course, not exclusive. The properties of an information network define its expressiveness and thus its ability to model data of a diverse nature, e.g. ontologies or experimental data.

2.1 Properties of Information Units

The basic information unit does not posses any additional semantical information. However, they will at least include a label attached to them in order to identify the object or concept they represent. Additional properties are the following:

Attributed. units of information can have additional attributes attached to them. An attribute might be a link to the original data it stems from, or a translation of a user-readable label. These attributes might be considered while reasoning or analyzing the network but do not carry general semantic information, such as the following properties.

Typed. information units carry an additional label that is used to distinguish between different semantics of information units, e.g. gene or protein. These types can additionally be organized in a hierarchy or an ontology.

Hierarchical. information units represent a sub-graph composed of any number of information units and relations that can be used to condense parts of the network or to represent more complex concepts such as cellular processes.

2.2 Properties of Relations

The basic connection between information units represents a relationship between the corresponding members. They are not required to carry a label.

Attributed. relations have attributes attached to them and also fall into this category. Similar to attributed information units, they can be considered during the reasoning process, but do not carry a general semantic information.

Typed. relations are similar to typed information units and can carry a label identifying their type. This attribute is used to distinguish between different semantics of relations such as activates or encodes. These types, as well as typed information units, can be organized in a hierarchy or an ontology.

Weighted. relations carry a special type of label - the weight - which represents the strength of a relation, e.g. a number reflecting the probability or strength of a correlation or some other measure of reliability that allows the integration of facts and pieces of evidence.

Directed. relations can be used to explicitly model relationships that are only valid in one direction, such as parent child dependency in a hierarchy.

Multi-relation. relations are generally represented as edges supporting only two members. Topic maps (see Section 3.3) in contrast represent relations as multi edges supporting any number of members. This allows a more flexible modeling of relationships with any number of members, e.g. co-expressed genes of an experiment or co-authors of a paper. Furthermore connections among relations themselves can be represented. Note that it is complicated to combine this property with the directed property mentioned above. Additional information would need to be provided, such as an embedding graph to identify sources and targets in a relation with more than two members.

3 Prominent Types of Information Networks

This section describes prominent types of information networks and characterizes them based on the previously discussed properties (see section 2) they support.

3.1 Ontologies

Ontologies are based on typed and directed relations using a controlled vocabulary for information units and relations dedicated to a certain domain. The creation of the curated vocabulary leads in general to a manual or semi-automatic creation of an ontology, requiring a comprehensive knowledge of the area to be described.

Figure 1 depicts a simple ontology where information units are represented as nodes and relations are represented as labeled arrows.

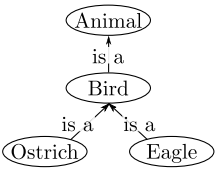


Fig. 1. Example of an ontology

In the area of life sciences particularly, many ontologies have been developed to share data from diverse research areas such as chemistry, biology or pharmacokinetics. One of the probably best known and most integrated ontologies in the biological field is the Gene Ontology (GO) [17]. The GO consists of three main ontologies describing the molecular function, biological process and cellular component of genes.

An attempt to integrate diverse ontologies has been made by the Open Biomedical Ontologies (OBO) consortium [52]. They have created a file exchange format and over 60 ontologies for different domains defining a general vocabulary that can be used by other systems.

A classification of biomedical ontologies has been completed by Bodenreider [10]. He classified these ontologies into three major categories: knowledge management; data integration, exchange and semantic interoperability; decision support and reasoning.

An ontology-based data integration platform is described in [33]. The authors describe a system that extends the existing text-mining framework ONDEX. ONDEX uses a core set of ontologies, which are aligned by several automated methods to integrate biological databases. The existing system is extended to support not only the alignment and integration of texts but heterogeneous data sources. The data is represented as a graph with attributed edges.

Tzitzikas et al. [56] describe a system that is based on the hierarchical integration of ontologies from different data sources. The system uses a mediator ontology, which bridges the heterogeneity of the different data source ontologies.

3.2 Semantic Networks

Semantic networks use typed relations to model the semantic of the integrated information units and their relations. Information units in semantic networks, in contrast to ontologies, are not represented by a curated vocabulary but rather described by attaching any number of attributes to them whose semantic is defined by the type of the relation.

Most of the semantic networks rely on Semantic Web [8] technologies such as the Resource Description Framework (RDF) [40], RDF Vocabulary Description Language (RDF Schema) and the Web Ontology Language (OWL) defined by the W3C consortium[2]. RDF is a knowledge representation and storage framework that uses triples. A triple consists of a subject, predicate and object. The subject and object are information units that are connected by a directed relation defined by the predicate.

In Figure 2 subjects and objects that are uniquely identifiable are depicted in ellipses, whereas objects containing values are depicted in boxes. Predicates are shown as arrows pointing from the object to the subject with the type of the relation as an annotation.

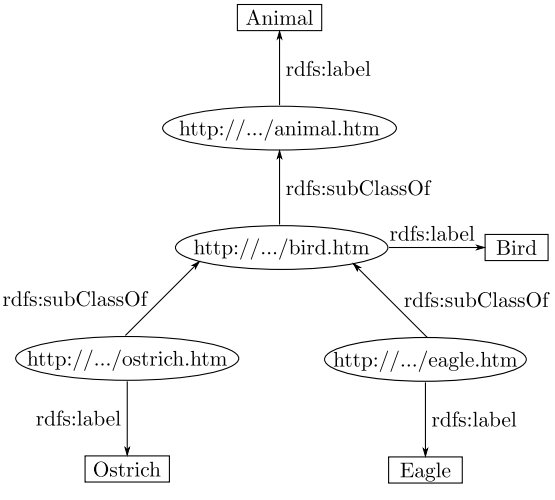


Fig. 2. Graph representation of a Semantic Web

The RDF Schema defines a core vocabulary that can be used to describe properties and classes. These properties and classes can be used to describe the members of a triple. OWL extends the RDF Schema by providing a set of additional standard terms to describe properties and classes in more detail such as relations between classes. It also defines the behavior of properties, e.g. symmetry or transitivity. OWL as well as the RDF Schema extend RDF by providing the means to model the semantics of the integrated data therefore enabling machines to make sense of the data. They are both described using the RDF.

[2] `http://www.w3.org/2001/sw/`

Bales and Johnson [5] analyzed large semantic networks created from 1998-2005 that involve both a graph theoretic perspective and semantic information. The results indicate that networks derived from natural language share common topological properties, such as scale-free and small-world characteristics.

Chen et al. [13] provide an introduction to semantic networks and semantic graph mining. In four case studies, they demonstrate the usage of semantic web technologies to analyze disease-causal genes, GO category cross-talks, drug efficacy and herb-drug interactions.

Belleau et al.[7] propose the Bio2RDF project to integrate data from different biological sources. Bio2RDF is used to integrate data from more than twenty different public bioinformatic sources by converting them into the RDF format.

YeastHub [15] another RDF-based data integration approach likewise integrates the data from heterogeneous sources into a RDF-based data warehouse. In addition they propose a standard RDF format for tabular data integration. The format can be used to convert any data table into a standardized RDF format.

A loosely coupled integration of semantic networks is proposed by Smith et al. [51] in the form of the LinkHub system. The system consists of smaller networks that can be connected by sharing a common hub. Thus independently maintained networks can be connected to the whole system by connecting them to one of the already integrated sub networks.

Biozon [9] combines the flexible graph structure with an ontology for vertex and edge types similar to the semantic web approach. This combined approach allows a more detailed description of a biological entity by either imposing more constraints on its nature in the hierarchy or on the structure of its relations to other entities in the graph. All vertices within Biozon are direct analogs to physical entities and sets of entities. Proteins, for example, are identified by their sequence of amino acids. In contrast to pure semantic networks Biozon allows any number of attributes to be attached to information units as well as to relations.

3.3 Topic Maps

Topic maps [23,47] use typed information units and relations. Furthermore topic maps support the modeling of multi relations with any number of members. The semantic of a topic is described by attaching any number of attributes to it.

Figure 3 depicts the three major elements of a topic map: topics (ellipses), associations (solid lines) and occurrences (boxes). Association and occurrence types are connected by the dashed lines whereas occurrences are connected by the dotted line.

A topic can generally be anything, for example a person, a concept or an idea. Topics can be assigned zero or more topic types, which are, in turn, defined as topics describing the semantics of the topic such as gene or protein.

Relations between any number of topics are represented by so-called associations. Associations are assigned a type that describes the association in more detail. Members of associations play a certain role defined by the association role. As with topic and occurrence types, association types and association roles

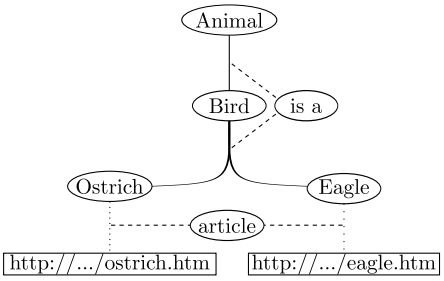


Fig. 3. Example of a topic map

are defined as topics themselves. In order to attach attributes to an association it needs to be converted into a topic by the act of reification.

Information resources that represent a topic or describe it in more detail are linked to topics by so-called occurrences. Occurrences are not generally stored in the topic map itself but are referenced using mechanisms supported by the system, e.g. Uniform Resource Identifiers (URI). Occurrences can have any number of different types, so-called occurrence types, that describe their semantics. These types are also defined as topics. Topic maps are self-documenting due to the fact that virtually everything in topic maps is a topic in the map itself, forming the ontology of the used topics and relation types.

An example of a topic-map-like data integration approach is PathSys [4]. In PathSys a relation is also represented as a vertex. This approach models relationships between relations themselves. To distinguish between information units and relations they introduce vertex types. Besides primary vertices representing information units and connector vertices representing relationships, they also introduce graph vertices. By introducing graph vertices, PathSys combines the multi relation property of topic maps with the hierarchical information unit property allowing the sub-graph representation to describe more complex objects such as protein complexes or cellular processes.

3.4 Weighted Networks

In most weighted networks the edge weight represents the strength of a relation such as reliability or probability. Weighted networks often exhibit additional properties such as types in order to be more expressive by modeling the semantic of the integrated data sources. They generally only support relationships with two members represented by the edges of the graph.

Figure 4 depicts a weighted network modeling the probability of a bird to be either a bird of prey or a flightless bird.

Probabilistic Weights. Probabilistic networks model the probability of the existence of a relationship. They are mostly used in the biological field to model interaction networks, e.g. gene-gene or protein-protein interaction networks. In order to model the probability of the relations the networks often depend

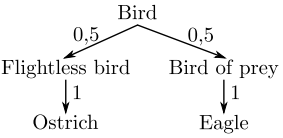


Fig. 4. Example of a weighted network

on a specific network structure or weight distribution. Bayesian networks, for example, depend on a directed acyclic graph, whose vertices model the random variables an its relations indicate their conditional dependencies [46].

Franke et al. [24] use three steps to fuse the information from the GO with microarray co-expression results and protein-protein interaction data using naive Bayesian networks. The resulting network called Genenetwork can be used to detect genes that are related to a disease based on genetic mutation.

Li et al. [41] use a two-layered approach to integrate gene relations from heterogeneous data sources. The first layer creates a fully connected Bayesian network for each integrated source, which represents the gene functional relations. The second layer combines these relations from the different data sources into one integrated network using a naive Bayesian method.

Jansen et al. [31] likewise propose a combination of naive Bayesian networks and fully connected Bayesian networks to create a protein-protein interaction network. They use the fully connected Bayesian networks to integrate experimental interaction data and naive Bayesian networks to incorporate other genomic features such as the biological process from the GO. To combine all results they use a naive Bayesian network as well.

In [55], Troyanskaya et al. introduce MAGIC (Multisource Association of Genes by Integration of Clusters). For each integrated data source, MAGIC creates a gene-gene relationship matrix to predict the functional relationship of two given genes. The matrices are generated from diverse high-throughput techniques such as gene expression microarrays. These gene-gene relationship matrices are weighted by the confidence in the integrated source and combined into a single matrix. This approach allows genes to be members of more than one group, which subsequently allows fuzzy clustering.

Heuristic Weights. Heuristic weights are mostly used to model the reliability or relevance of a given relation, thus allowing the integration of well-curated sources such as ontologies and pieces of evidence such as noisy experimental data in a single network.

In order to integrate data from diverse biological sources for protein function prediction, Chua et al. [16] propose Integrated Weighted Averaging (IWA). This combines local prediction methods with a global weighting strategy. Each data source is transformed into an undirected graph with proteins as vertices and relationships between proteins as edges. Each source graph has a score reflecting its reliability. Finally, all source graphs are combined in a single graph using IWA.

Kiemer et al. [32] use a weighted network to integrate yeast protein information from different data sources forming a protein-protein interaction network

called WI-PHI. The network consists of 50,000 interactions from all data sources. The edge weight of the WI-PHI network is computed using the socio-affinity index [26], quantifying the propensity of proteins to form partnerships, multiplied by a weight constant per integrated data source defining its accuracy.

In Biomine [49] the edge weight is a combination of three different weights: reliability, relevance and rarity. Reliability reflects the reliability of the source the edge stems from. By changing the relevance of different node or edge types, e.g. proteins, genes, a user can focus on the types he or she is most interested in. Finally rarity is computed using the degree of the incident vertices. Edges that connect vertices with a low degree have a higher rarity score than edges that connect vertices with a high degree. Vertices and edges have a type assigned describing their nature. Each edge has its inverse edge with a natural inverse type such as "coded by" and "is referred by". Thus forming a weighted undirected graph with directed edge types.

In the next section we describe bisociative information networks that combine the properties of the existing network types in order to support the integration of heterogeneous data sources.

4 BisoNets: Bisociative Information Networks

Bisociative information networks (BisoNets) provide the flexibility to integrate relations from semantically meaningful information as well as loosely coupled information fragments with any number of members by adopting a weighted k-partite graph structure (see Figure 5).

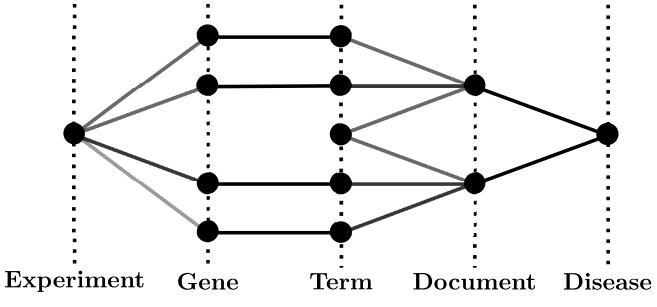


Fig. 5. Example of a 5-partite BisoNet

Vertices in BisoNets represent arbitrary units of information, e.g., a gene, protein, specific molecule, index term, or document, or abstract concepts such as ideas, acts or events. Vertices of the same type are grouped into vertex partitions such as documents, authors, genes or experiments. Since a vertex can play diverse roles it can be assigned to several partitions.

Depending on a certain view, the vertices of a partition can act as relations or information units. Let us consider a document author network to illustrate this

concept. In one view the documents can describe the relationship between co-authors. Whereas in another view the authors describe the relationship between documents that have been written by the same authors. Thus the role of a vertex partition depends on the current view on the data.

Connections between vertices are represented by edges. An edge can only exist between vertices of diverse partitions; this leads to the k-partite graph structure. Hence a relation between two information units (e.g., authors) is described by a third information unit (e.g., document). A BisoNet therefore consists of at least two partitions, the first partition representing the information units and the second partition describing the relations between the information units.

The certainty of a connection is represented by the weight of the edge. A stronger weight represents a higher certainty in the existence of the connection. Thus, a connection derived from a reliable data source (e.g., a manually curated ontology) is assigned a stronger weight than a connection derived from an automated method (e.g., text mining method).

BisoNets model the main characteristics of the integrated information repositories without storing all the detailed data from which these characteristics are derived. By focusing on the concepts and their relations alone, BisoNets therefore allow very large amounts of data to be integrated.

Definition 1 (BisoNet). *A BisoNet* $B = (V_1, ..., V_k, E, \lambda, \omega)$ *is an attributed graph, where* $V = \bigcup_{i \leq k} V_i$ *represents the union of all vertex partitions and* $k \geq 2$ *denotes the number of existing partitions. Every vertex* $v \in V$ *represents a unit of information and can be a member of multiple partitions.*

The set of edges $E = \{\{u, v\} : u \in V_i; v \in V_j; j \neq i\}$ *connects vertices from two different vertex partitions, whereas an edge* $e = \{u, v\} \in E$ *represents a connection between the two vertices* $u \in V_i$ *and* $v \in V_j$ *where* $i \neq j$ *and* $2 \leq i, j \leq k$.

The function $\lambda : V \to \Sigma^*$ *assigns each vertex* $v \in V$ *an unique label from* Σ^*. *This allows for the identification of a vertex by its unique label.*

The certainty of a relation is represented by the weight of an edge $e \in E$, *which is assigned by the function* $\omega : E \to [0, 1]$ *and where a weight of 1 represents the highest certainty.*

4.1 Summary

Table 1 compares the prominent types of information networks from section 3 with BisoNets based on the properties they support. The table shows that most of the networks support typed relations whereas topic maps and BisoNets also support typed information units. The types enable us to distinguish between different types of information units and relations, leading to to a better understanding of the integrated data. In addition the type information allows semantical information to be processed by a computer system. But the usage of type information requires detailed knowledge about the information that should be integrated into the network. The creation of a suitable type collection that allows the integration of data from diverse sources is thus an elaborated task which

Table 1. Properties matrix of prominent types of information network in conjunction with BisoNets (A=Attributed, T=Typed, H=Hierarchical, W=Weighted, D=Directed and M=Multi relation)

	Information Units			Relations				
	A	T	H	A	T	W	D	M
Ontologies					X		X	
Semantic Networks	X				X		X	
Topic Maps	X	X		X	X			X
Weighted Networks						X		
BisoNets	X	X	X	X	X	X	X	X

often has to be done manually. Moreover, not all data sources do possess the required semantical information to assign the right type and therefore manual annotations of the integrated information units and relations might be required. If information units and relation types are abandoned, the integration of data from heterogeneous sources is much easier but it might make the comprehension of the integrated data more difficult. As a result, BisoNets support typed information units and relations and allow their usage if the integrated data sources provide this information, however they are not mandatory. In contrast to topic maps, BisoNets also support weighted relations, thus allowing not only the integration of facts but also pieces of evidence. BisoNets combine the properties of the existing network types in order to provide a well-defined and powerful data structure that provides the flexibility to integrate relations from heterogeneous data sources.

5 Patterns of Bisociation in BisoNets

Once the information has been integrated into a BisoNet, it can be analyzed in order to find interesting patterns in the integrated data. One class of pattern is bisociation. So far, we have identified three different kinds of bisociations [37], which are described in more detail below.

5.1 Bridging Concept

Bridging concepts connect dense sub-graphs from different domains (see Figure 6). Bridging concepts employ ambiguous concepts or metaphors and are often used in humor [34] and riddles [19]. While ambiguity is useful for making jokes or telling stories, it is less popular in serious scientific or engineering applications. For example, the concept of a “jaguar” is ambiguous since it may refer to either an animal or a car. Metaphors, on the other hand, describe a form of understanding or reasoning in which a concept or idea in one domain is understood or viewed in terms of concepts or ideas from another domain. The statement “You are wasting my time”, for instance, can be seen as a metaphor that connects the time with the financial domain. Metaphors play a major role in

our everyday life as they afford a degree of flexibility that facilitates discoveries by connecting seemingly unrelated subjects [39].

A first approach to detect bridging concepts is the discovery of concept graphs [35,36] in the integrated data. Concept graphs can be used to identify existing and missing concepts in a network by searching for densely connected quasi bi-partite sub-graphs. Once a concept graph has been detected the domains, its aspect and member vertices stem from, can be analyzed in order to find concepts graphs, e.g. concepts that connect information units from different domains.

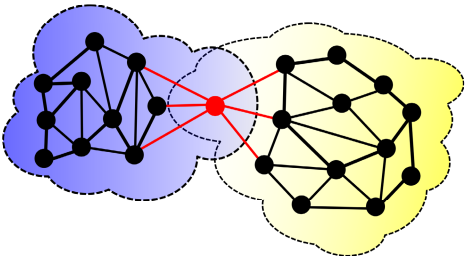
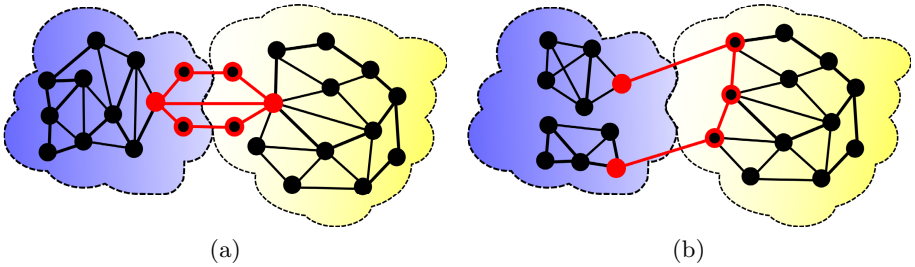


Fig. 6. Bridging concept

5.2 Bridging Graphs

Bridging graphs are sub-graphs that connect concepts from different domains (see Figure 7). They may lead to surprising information arising from different domains since they are able to link seemingly unrelated domains (see Figure 7a). An example of where bridging graph could be used to realize bisociation is the Eureka act of the Archimedes example [20]. A bridging graph may also lead to the linking of two disconnected concepts from the same domain via a connection through and unrelated domain (see Figure 7b).

A first step in the direction of the discovery of bridging graphs is the formalization and detection of such domain-crossing sub-graphs [43,44]. The discovered sub-graphs can be further ranked according to their potential interestingness. Therefore the interestingness is measured by a so called b-score that takes into account the size of the connected domains, the sparsity of the connections between the different domains and the distribution of the neighbors of the bridging vertices.



(a) (b)

Fig. 7. Bridging graphs

5.3 Bridging by Graph Similarity

Bisociations based on graph similarity are represented by sub-graphs of two different domains that are structurally similar (see Figure 8). This is the most abstract pattern of bisociation that has the potential to lead to new discoveries by linking domains that do not have any connection except for the similar interaction of the bridging concepts and their neighbors.

These structurally similar but disconnected regions in a BisoNet can be discovered by means of a vertex similarity based on the structural properties of vertices. In [53,54] a spatial similarity (activation similarity) and a structural similarity (signature similarity) based on spreading activation are introduced, which can be used in combination in order to identify bisociations based on structurally similar but disconnected sub-graphs.

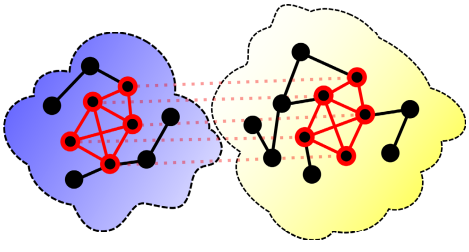


Fig. 8. Bridging by graph similarity

6 Conclusion

In this chapter we identified several key properties of information units and relations used in information networks. We provided an overview of different types of information networks and categorized them based on the identified properties. These properties reflect the expressiveness and thus the ability of an information network to model data of a diverse nature.

We further describe BisoNets as a new type of information network that is tailored to the integration of heterogeneous data sources from diverse domains. They possess the main properties required to integrate large amounts of data from a variety of information sources. By supporting weighted edges BisoNets support the integration not only of facts such as hand curated ontologies but also of pieces of evidence such as results from biological experiments.

Finally we described three patterns of bisociations in BisoNets. Bridging concepts refer to a single vertex that is connected to vertices from different domains. These vertices, which belong to multiple domains, might be an indication of ambiguity or metaphor - metaphors often being used in humor and riddles. Bridging graphs on the other hand are sub-graphs consisting of multiple vertices and edges that connect concepts from different domains. These sub-graphs might lead to new insights by connecting seemingly unrelated domains. Last but not least, domain bridging by structural similarity is the most abstract pattern of bisociation with the potential to lead to truly new discoveries by linking domains that are

otherwise unconnected, except for the similar structure of their corresponding sub-graphs.

References

1. Abello, J., Korn, J.: Mgv: a system for visualizing massive multidigraphs. Transactions on Visualization and Computer Graphics 8(1), 21–38 (2002)
2. Albert, R., Barabasi, A.-L.: Statistical mechanics of complex networks. Reviews of Modern Physics 74, 47–97 (2002)
3. Auillans, P., de Mendez, P.O., Rosenstiehl, P., Vatant, B.: A Formal Model for Topic Maps. In: Horrocks, I., Hendler, J. (eds.) ISWC 2002. LNCS, vol. 2342, pp. 69–83. Springer, Heidelberg (2002)
4. Baitaluk, M., Qian, X., Godbole, S., Raval, A., Ray, A., Gupta, A.: Pathsys: integrating molecular interaction graphs for systems biology. BMC Bioinformatics 7, 55 (2006)
5. Bales, M.E., Johnson, S.B.: Graph theoretic modeling of large-scale semantic networks. Journal of Biomedical Informatics 39, 451–464 (2006)
6. Belew, R.: Adaptive information retrieval: using a connectionist representation to retrieve and learn about documents. In: Proceedings of the 12th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 11–20 (1989)
7. Belleau, F., Nolin, M.-A., Tourigny, N., Rigault, P., Morissette, J.: Bio2rdf: towards a mashup to build bioinformatics knowledge systems. Journal of Biomedical Informatics 41, 706–716 (2008)
8. Berners-Lee, T., Hendler, J., Lassila, O.: The semantic web. Scientific American 5, 34–43 (2001)
9. Birkland, A., Yona, G.: Biozon: a system for unification, management and analysis of heterogeneous biological data. BMC Bioinformatics 7, 70 (2006)
10. Bodenreider, O.: Biomedical ontologies in action: role in knowledge management, data integration and decision support. IMIA Yearbook of Medical Informatics 1, 67–79 (2008)
11. Brandes, U., Erlebach, T.: Network Analysis: Methodological Foundations. Springer (2005)
12. Burgun, A., Bodenreider, O.: Accessing and integrating data and knowledge for biomedical research. IMIA Yearbook of Medical Informatics 1, 91–101 (2008)
13. Chen, H., Ding, L., Wu, Z., Yu, T., Dhanapalan, L., Chen, J.Y.: Semantic web for integrated network analysis in biomedicine. Briefings in Bioinformatics 10, 177–192 (2009)
14. Chen, H., Ng, T.: An algorithmic approach to concept exploration in a large knowledge network (automatic thesaurus consultation): Symbolic branch-and-bound search vs. connectionist hopfield net activation. Journal of the American Society for Information Science 46(5), 348–369 (1995)
15. Cheung, K.-H., Yip, K.Y., Smith, A., Deknikker, R., Masiar, A., Gerstein, M.: Yeasthub: a semantic web use case for integrating data in the life sciences domain. Bioinformatics 21(suppl.1), i85–i96 (2005)

16. Chua, H.N., Sung, W.-K., Wong, L.: An efficient strategy for extensive integration of diverse biological data for protein function prediction. Bioinformatics 23, 3364–3373 (2007)
17. Consortium, G.O.: Creating the gene ontology resource: design and implementation. Genome Research 11, 1425–1433 (2001)
18. Crestani, F.: Application of spreading activation techniques in information retrieval. Artificial Intelligence Review 11, 453–482, 12 (1997)
19. Dienhart, J.M.: A linguistic look at riddles. Journal of Pragmatics 31(1), 95–125 (1999)
20. Dubitzky, W., Kötter, T., Schmidt, O., Berthold, M.R.: Towards Creative Information Exploration Based on Koestler's Concept of Bisociation. In: Berthold, M.R. (ed.) Bisociative Knowledge Discovery. LNCS (LNAI), vol. 7250, pp. 11–32. Springer, Heidelberg (2012)
21. Durand, P., Labarre, L., Meil, A., Divol, J.-L., Vandenbrouck, Y., Viari, A., Wojcik, J.: Genolink: a graph-based querying and browsing system for investigating the function of genes and proteins. BMC Bioinformatics 7(1), 21 (2006)
22. Figeys, D.: Combining different 'omics' technologies to map and validate protein-protein interactions in humans. Briefings in Functional Genomics and Proteomics 2, 357–365 (2004)
23. I.O. for Standardization. Information Technology – Document Description and Processing Languages – Topic Maps – Data Model. ISO, Geneva, Switzerland (2006)
24. Franke, L., van Bakel, H., Fokkens, L., de Jong, E.D., Egmont-Petersen, M., Wijmenga, C.: Reconstruction of a functional human gene network, with an application for prioritizing positional candidate genes. The American Journal of Human Genetics 78, 1011–1025 (2006)
25. Furnas, G.W.: Generalized fisheye views. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, vol. 17(4), pp. 16–23 (1986)
26. Gavin, A.-C., Aloy, P., Grandi, P., Krause, R., Boesche, M., Marzioch, M., Rau, C., Jensen, L.J., Bastuck, S., Dümpelfeld, B., Edelmann, A., Heurtier, M.-A., Hoffman, V., Hoefert, C., Klein, K., Hudak, M., Michon, A.-M., Schelder, M., Schirle, M., Remor, M., Rudi, T., Hooper, S., Bauer, A., Bouwmeester, T., Casari, G., Drewes, G., Neubauer, G., Rick, J.M., Kuster, B., Bork, P., Russell, R.B., Superti-Furga, G.: Proteome survey reveals modularity of the yeast cell machinery. Nature 440, 631–636 (2006)
27. Getoor, L., Diehl, C.: Link mining: a survey. ACM SIGKDD Explorations Newsletter 7(2), 3–12 (2005)
28. Han, J.: Mining Heterogeneous Information Networks by Exploring the Power of Links. In: Gama, J., Costa, V.S., Jorge, A.M., Brazdil, P.B. (eds.) DS 2009. LNCS, vol. 5808, pp. 13–30. Springer, Heidelberg (2009)
29. Hayes, J.: A graph model for RDF. Master's thesis, Technische Universität Darmstadt, Dept. of Computer Science, Darmstadt, Germany. In: Collaboration with the Computer Science Dept., University of Chile, Santiago de Chile (2004)
30. Hayes, J., Gutierrez, C.: Bipartite Graphs as Intermediate Model for RDF. In: McIlraith, S.A., Plexousakis, D., van Harmelen, F. (eds.) ISWC 2004. LNCS, vol. 3298, pp. 47–61. Springer, Heidelberg (2004)
31. Jansen, R., Yu, H., Greenbaum, D., Kluger, Y., Krogan, N.J., Chung, S., Emili, A., Snyder, M., Greenblatt, J.F., Gerstein, M.: A bayesian networks approach for predicting protein-protein interactions from genomic data. Science 302, 449–453 (2003)
32. Kiemer, L., Costa, S., Ueffing, M., Cesareni, G.: Wi-phi: A weighted yeast interactome enriched for direct physical interactions. Proteomics 7, 932–943 (2007)

33. Koehler, J., Rawlings, C., Verrier, P., Mitchell, R., Skusa, A., Ruegg, A., Philippi, S.: Linking experimental results, biological networks and sequence analysis methods using ontologies and generalised data structures. Silico Biology 5, 33–44 (2005)
34. Koestler, A.: The Act of Creation. Macmillan (1964)
35. Kötter, T., Berthold, M.R.: (Missing) concept discovery in heterogeneous information networks. In: Proceedings of the 2nd International Conference on Computational Creativity, pp. 135–140 (2011)
36. Kötter, T., Berthold, M.R.: (Missing) Concept Discovery in Heterogeneous Information Networks. In: Berthold, M.R. (ed.) Bisociative Knowledge Discovery. LNCS (LNAI), vol. 7250, pp. 230–245. Springer, Heidelberg (2012)
37. Kötter, T., Thiel, K., Berthold, M.R.: Domain bridging associations support creativity. In: Proceedings of the International Conference on Computational Creativity, pp. 200–204 (2010)
38. Kwoh, C.K., Ng, P.Y.: Network analysis approach for biology. Cellular and Molecular Life Sciences 64, 1739–1751 (2007)
39. Lakoff, G., Johnson, M.: Metaphors We Live by. University of Chicago Press (1980)
40. Lassila, O., Swick, R.R.: Resource Description Framework (RDF) model and syntax specification. W3C Working Draft (February 2002)
41. Li, J., Li, X., Su, H., Chen, H., Galbraith, D.W.: A framework of integrating gene relations from heterogeneous data sources: an experiment on arabidopsis thaliana. Bioinformatics 22(16), 2037–2043 (2006)
42. Martinez Morales, A.A.: A directed hypergraph model for RDF. In: Simperl, E., Diederich, J., Schreiber, G. (eds.) Proceedings of the KWEPSY 2007, vol. 275 (2007)
43. Nagel, U., Thiel, K., Kötter, T., Piątek, D., Berthold, M.R.: Bisociative Discovery of Interesting Relations between Domains. In: Gama, J., Bradley, E., Hollmén, J. (eds.) IDA 2011. LNCS, vol. 7014, pp. 306–317. Springer, Heidelberg (2011)
44. Nagel, U., Thiel, K., Kötter, T., Piatek, D., Berthold, M.R.: Towards Discovery of Subgraph Bisociations. In: Berthold, M.R. (ed.) Bisociative Knowledge Discovery. LNCS (LNAI), vol. 7250, pp. 263–284. Springer, Heidelberg (2012)
45. Pavlopoulos, G., Wegener, A.-L., Schneider, R.: A survey of visualization tools for biological network analysis. BioData Mining 1(1), 1–12 (2008)
46. Pearl, J.: Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference. Morgan Kaufmann Publishers (1988)
47. Pepper, S.: The tao of topic maps: finding the way in the age of infoglut. In: Proceedings of XML Europe (2000)
48. Schaeffer, S.E.: Graph clustering. Computer Science Review 1, 27–64 (2007)
49. Sevon, P., Eronen, L., Hintsanen, P., Kulovesi, K., Toivonen, H.: Link Discovery in Graphs Derived from Biological Databases. In: Leser, U., Naumann, F., Eckman, B. (eds.) DILS 2006. LNCS (LNBI), vol. 4075, pp. 35–49. Springer, Heidelberg (2006)
50. Shen, Z., Ma, K.-L., Eliassi-Rad, T.: Visual analysis of large heterogeneous social networks by semantic and structural abstraction. IEEE Transactions on Visualization and Computer Graphics 12(6), 1427–1439 (2006)
51. Smith, A.K., Cheung, K.-H., Yip, K.Y., Schultz, M., Gerstein, M.K.: Linkhub: a semantic web system that facilitates cross-database queries and information retrieval in proteomics. BMC Bioinformatics 8, S5 (2007)
52. Smith, B., Ashburner, M., Rosse, C., Bard, J., Bug, W., Ceusters, W., Goldberg, L.J., Eilbeck, K., Ireland, A., Mungall, C.J., Consortium, O.B.I., Leontis, N., Rocca-Serra, P., Ruttenberg, A., Sansone, S.-A., Scheuermann, R.H., Shah, N.,

Whetzel, P.L., Lewis, S.: The obo foundry: coordinated evolution of ontologies to support biomedical data integration. Nature Biotechnology 25, 1251–1255 (2007)
53. Thiel, K., Berthold, M.R.: Node similarities from spreading activation. In: Proceedings of the IEEE International Conference on Data Mining (2010)
54. Thiel, K., Berthold, M.R.: Node Similarities from Spreading Activation. In: Berthold, M.R. (ed.) Bisociative Knowledge Discovery. LNCS (LNAI), vol. 7250, pp. 246–262. Springer, Heidelberg (2012)
55. Troyanskaya, O.G., Dolinski, K., Owen, A.B., Altman, R.B., Botstein, D.: A bayesian framework for combining heterogeneous data sources for gene function prediction (in saccharomyces cerevisiae). Proceedings of the National Academy of Sciences 100, 8348–8353 (2003)
56. Tzitzikas, Y., Constantopoulos, P., Spyratos, N.: Mediators over ontology-based information sources. In: Second International Conference on Web Information Systems Engineering, pp. 31–40 (2001)
57. van Ham, F., van Wijk, J.: Interactive visualization of small world graphs. In: van Wijk, J. (ed.) Proc. IEEE Symposium on Information Visualization INFOVIS 2004, pp. 199–206 (2004)