Erschienen in: Bisociative Knowledge Discovery : An Introduction to Concept, Algorithms, Tools, and Applications / Michael R. Berthold (ed.). - Heidelberg [u.a.] : Springer, 2012. - S. 263-284. - (Lecture Notes in Computer Science ; 7250 : Lecture notes in artificial intelligence). - ISBN 978-3-642-31829-0

Towards Discovery of Subgraph Bisociations

Uwe Nagel*, Kilian Thiel, Tobias Kötter,
Dawid Piątek, and Michael R. Berthold

Nycomed-Chair for Bioinformatics and Information Mining
Dept. of Computer and Information Science
University of Konstanz
firstname.lastname@uni-konstanz.de

Abstract. The discovery of surprising relations in large, heterogeneous information repositories is gaining increasing importance in real world data analysis. If these repositories come from diverse origins, forming different domains, domain bridging associations between otherwise weakly connected domains can provide insights into the data that are not accomplished by aggregative approaches. In this paper, we propose a first formalization for the detection of such potentially interesting, domain-crossing relations based purely on structural properties of a relational knowledge description.

1 Motivation

Classical data mining approaches propose (among others) two major alternatives to exploit data collections. One scenario tries to fit a model to the given data and thereby to predict the behavior of some underlying system. Another approach describes all or part of the given data by patterns such as clusters or frequent itemsets to provide an insight or highlight mechanisms that led to such patterns. Both variants have in common that some hypothesis about the considered data is involved and that the processing is motivated by a concrete question. A necessity for such a motivated processing is some a priori knowledge or decision about either the involved data (i.e. what type of model could be fitted) or the form in which findings are described (e.g. clusters or frequent item sets). While in the first case the possible findings are narrowed by the aspect of the systems behavior that is modelled, in the latter case the choice of patterns limits the possible findings. In short, one could say that in those approaches the problem to be solved is simplified by narrowing it down through the investment of a priori knowledge or by specifying the form of outcome.

Alternatively, Explorative (or Visual) Data Mining attempts to overcome this problem by creating a more abstract overview of the entire data together with subsequent drill-down operations. Thereby it additionally enables the search for arbitrary interesting patterns on a structural level, detached from the semantics of the represented information. However, such overviews still leave the entire

* Corresponding author.

M.R. Berthold (Ed.): Bisociative Knowledge Discovery, LNAI 7250, pp. 263–284, 2012.

search for interesting patterns to the user and therefore often fail to actually point to interesting and truly novel details.

We propose a different approach: instead of considering all available data or large parts of it at once, we concentrate on the identification of interesting, seldom appearing details. In that, integrated data is explored by finding unexpected and potentially interesting connections that hopefully trigger the user's interest, ultimately supporting creativity and outside-the-box thinking. Our assumption is that such connections qualify as unexpected by connecting seemingly unrelated domains. As already pointed out by Henri Poincaré [14]: "Among chosen combinations the most fertile will often be those formed of elements drawn from domains which are far apart... Most combinations so formed would be entirely sterile; but certain among them, very rare, are the most fruitful of all." A historical example for such a combination is provided by the theory of electromagnetism by Maxwell [11], connecting electricity and magnetism. Consequently, we embrace the diversity of different data sources and domains of knowledge. We thus do not fuse those into one large homogeneous knowledge base by e.g. mapping them into a common feature space. Instead, we model the given data sparsely, try to identify (possibly hidden) domains and search for rare instead of frequent and weak instead of strong patterns, i.e. exclusive, domain crossing connections. Though for technical integration we still need a certain homogeneity in the data representation.

With respect to this demand, we assume a knowledge representation fulfilling only very few conditions: the representation of *information units* and links between them without any further attributes. Based on that, we address two sub-problems: the identification of domains and the assessment of the potential interestingness of connections between these domains.

2 Networks, Domains and Bisociations

In this section, we transfer the theoretical concept of domain crossing associations which are called *bisociations* [10] (to emphasize the difference to associations within a single domain) to a setting where a relational description of knowledge is given. We will explain the model that incorporates our knowledge base, narrow down the concepts underlying *domains* and *bisociations* and identify properties that allow to assess the interestingness of a bisociation.

2.1 Knowledge Modeling

As a preliminary, we assume that the available knowledge is integrated into a unifying data model. We model this as an undirected, unweighted graph structure with nodes representing units of information and edges representing their relations. Examples for information units are terms, documents, genes or experiments. Relations could arise from references, co-occurrences or explicitly encoded expert knowledge. The only semantic assumption we do make is, that the relation expressed by the links is of positive nature, i.e. is to be interpreted as similarity not dissimilarity.

A graph is described as $G = (V, E)$ with node set V, edge set $E \subseteq \binom{V}{2}$ and $n = |V|$ the number of nodes. The degree of a node, i.e. the number of incident edges, is denoted as $d(v)$ and its neighboring nodes in G as $N(v)$. We further access the structure of G via its adjacency matrix A, with $(A)_{uv} = 1$ if $\{u, v\} \in E$ and 0 otherwise. Finally, for a set of nodes $U \subseteq V$, $G[U]$ denotes the subgraph of G induced by the nodes of U, i.e. the nodes of U and all edges of E that connect two nodes of U.

The presented model does not contain any hints on the semantics behind the incorporated units of information except for their relations. In practice, such semantics would be provided by additional attributes attached to nodes and links. Our approach will, however, not employ them in any automatic processing, thereby ensuring maximal flexibility. Yet supporting attributes are helpful and necessary in the process of manual result interpretation and should consequently not be removed. In contrast they should be preserved completely: in the phase of manual interpretation they provide the necessary link to the semantic layer of the considered data. The fact that they are not employed in any automatic processing rules out the demand for homogeneity and thereby the necessity to convert them into a common format, leaving freedom to attach arbitrary information that might be useful.

While we ignore additional information attached to nodes and details about link interpretations, the general interpretation of links is an important aspect since the structural information provided by them is the sole basis of reasoning about the connected objects. In general, we consider different types of links expressing different relations and we allow some inhomogeneity within the set of links, such as different sources that led to the formation of links. However, we assume that within an individual data set all links obey roughly the same interpretation. Consider as an example a knowledge collection consisting of scientific articles. We would interpret the articles as information units and therefore model them as nodes. For the derivation of links we can choose between alternative semantic relations. As an example we could derive similarities by text analysis and derive links from these. Alternatively, we could exploit the fact that scientific articles reference each other and introduce a link for each reference. Both approaches have their assets and drawbacks and surely do not represent all possibilities of link derivation. In the identification of domains as described in the following, these two interpretations have to be handled differently which makes their distinction very important. We do not fix any decision about the type of links, but consider it as an important design decision in the process of data modeling and stress that it has to be considered carefully in the whole process. In the remainder we restrict our considerations to the two interpretations described above and point out where the method has to be adapted to the type of link interpretation.

2.2 Domains

As indicated before, in this context a *domain* is a set of information units from the same field or area of knowledge. Domains exist with different granularity and

thus can be partially ordered in a hierarchical way from specific to general. As an example consider the domains chemistry, biology, biochemistry and the science domain in general. While the first three are clearly subsets of the science domain, biochemistry is neither a proper subset of biology nor chemistry but overlaps with both of them. Furthermore, this distinction of domains may be sufficient in the context of common knowledge while scientists working in these fields would surely subdivide them. Consequently, the granularity of a domain depends on a specific point of view, which can be a very local one. In addition, information units may belong to several domains which are not necessarily related. The eagle for example belongs to the domain of animals and in addition to the coat of arms domain.

Relation to Graph Structure. Intuitively, a set of highly interconnected nodes indicates an intense interrelation that should be interpreted as a common domain. While this is a sound assumption when connections express similarities between the involved concepts, it is not necessarily true when links express other semantic relations. In the example of scientific articles a collection of papers approaching a common problem would signify an example domain. Yet the similarity of these articles is not necessarily reflected by mutual references, especially if they were written at the same time. However, they will very likely share a number of references. Consequently, we derive domains from common neighborhoods instead of relying on direct connections between information units. This allows domains to be identified when the connections express either references or similarities since nodes in a densely connected region also have similar neighborhoods. Two information units that share all (or - more realistically - almost all) their connections to other information units should therefore belong to a common domain. Since they are in this respect indistinguishable and their relations form the sole basis for our reasoning about them, all possibly identifiable domains have to contain either both or none of them. We will show a concrete node similarity that expresses this property and relaxes the conditions in Section 3. This similarity will then be used as a guidance in the identification of domains.

As mentioned before, the discussed guidelines are necessarily tailored to the considered link semantics and have to be adapted if links in the graph representation are derived differently. The interface between link interpretation and the process of domain identification is here established by a node similarity hinting at common domain affiliation. For the adaption to different link interpretations it is thus only necessary to adapt the node similarity correspondingly, ensuring that highly similar nodes tend to belong to identical domains on all levels while decreasing similarity indicates that nodes share fewer and thus only upper level domains.

Domain Recovery. Assuming a node similarity with the described properties, recursive merging of the most similar nodes leads to a merge tree as produced by hierarchical clustering. In the following, we consider the inner nodes of such a merge tree as candidates for domains.

The resulting domains form a hierarchy on the information units which is similar to an ontology but is not able to render all possible domain assignments. That is, any two domains resulting from this process overlap completely (one is contained in the other) or are completely disjoint. A number of domains could remain unidentified since cases of partially overlapping domains are excluded by the procedure, as the domain of biochemistry from the example above, given that biology and chemistry are identified as domains. We consider this as an unavoidable approximation for now, posing the extraction of domains as a separate problem.

2.3 Bisociations

A connection - usually indirect - between information units from multiple, otherwise unrelated domains is called *bisociation* in contrast to associations that connect information units within the same domain. The term was introduced by Koestler [9] in a theory to describe the creative act in humor, science and art.

Up to now, three different patterns of bisociation have been described in this context [10]: bridging concepts, bridging graphs and structural similarity. Here we focus on the discovery of bridging graphs, i.e. a collection of information units and connections providing a "bisociative" relation between diverse domains.

Among the arbitrary bisociations one might find, not all are going to be interesting. To assess their interestingness, we follow Boden [2] defining a creative idea in general as *new*, *surprising*, and *valuable*. All three criteria depend on a specific reference point: A connection between two domains might be long known to some specialists but new, surprising, and hopefully valuable to a specific observer, who is not as familiar with the topic. To account for this, Boden [2] defines two types of creativity namely H-creativity and P-creativity. While H-creativity describes globally (historical) new ideas, P-creativity (psychological) limits the demand of novelty to a specific observer. Our findings are most likely to be P-creative since the found connections have to be indicated by the analyzed data in advance. However, a novel combination of information sources could even lead to H-creative bisociations. Analogous to novelty, the value of identified bisociations is a semantically determined property and strongly depends on the viewers' perspective. Since both novelty and value cannot be judged automatically, we leave their evaluation to the observer. In contrast, the potential surprise of a bisociation can be interpreted as the unlikeliness of a connection between the corresponding domains. We will express this intuition in more formal terms and use it as a guideline for an initial evaluation of possible bisociations.

Identifying Bisociations. Based on these considerations, we now narrow down properties that are in our view essential for two connected domains to form a bisociation. Despite the discussion above, we have not yet given a technical definition of what a domain is. We will return to this problem in Section 3.1 and assume for now that a domain is simply a set of information units. In the graph representation, two domains are connected either directly by edges between their

nodes or more generally by nodes that are connected to both domains - the *bridging nodes*. Analogous to these more or less direct connections, of course connections spanning larger distances in the graph are possibly of interest. However, for the simplicity of description and due to the complexity of their determination, we will reduce the following considerations to the described simple cases. Such connecting nodes or edges bridge the two domains and together with the connected domains they form a *bisociation candidate*:

Definition 1 (Bisociation Candidate). *A* bisociation candidate *is a set of two domains and their connection within the network. That is, the subgraph induced by the nodes of the two domains* δ_1, δ_2 *and any further nodes that are connected to both domains:*

$$G\left[\delta_1 \cup \delta_2 \cup \{v \in V : N(v) \cap \delta_1 \neq \emptyset \wedge N(v) \cap \delta_2 \neq \emptyset\}\right]$$

Since it is impossible to precisely define what a surprising bisociation is, we develop three properties that distinguish promising bisociation candidates: *exclusiveness*, *size*, and *balance*. These technical demands are derived from an information-scientific view as e.g. expressed in [6]. In Ford's view, the creativity of a connection between two domains is related to (i) the dissimilarity of the connected domains and (ii) the level of abstraction on which the connection is established. In the following, we transfer these notions into graph theoretic terms by capturing them in properties that relate to structural features which can be identified in our data model.

We begin with the dissimilarity of two domains which we interpret as their connectedness within the graph. Maximal dissimilarity is rendered by two completely unconnected domains, closely followed by "minimally connected" domains. While the former case does not yield a bridging graph based bisociation (i.e. the connection itself is missing) the latter is captured by the exclusiveness property. Exclusiveness of a bisociation candidate states that the two involved domains are only sparsely connected, thereby expressing the demand of dissimilarity. On a more technical level it also excludes merely local exclusivity caused by nodes of high degree which connect almost everything, even unrelated domains, without providing meaningful connections.

Property (Exclusiveness)
A bisociation candidate is exclusive iff the connection between the two domains is

1. *small: the number of nodes connected to both domains (bridging nodes) is small in relation to the number of nodes in the adjacent domains;*
2. *sparse: the number of links between either the two domains directly or the domains and the nodes connecting them is small compared to the theoretical maximum;*
3. *concentrated: neighbors of the bridging nodes are concentrated in the adjacent domains and not scattered throughout the rest of the graph.*

Alternatively, this could be described in terms of probabilities: In a bisociation candidate two nodes from different domains are *linked* when they share an edge or have a common neighbor. Then exclusiveness describes the fact that two such nodes, randomly chosen from the two domains, are linked only with a low probability.

Directly entangled with this argument is the demand for size: a connection consisting of only a few nodes and links becomes less probable with growing domain sizes. In addition, a relation between two very small domains is hard to judge without knowledge about the represented semantic. It could be an expression of their close relation being exclusive only due to the small size of the connected domains. In that case the larger domains containing these two would show even more relations. It could also be an exclusive link due to domain dissimilarity. However, this situation would in turn be revealed when considering the larger domains, since these would also be exclusively connected. In essence, the exclusiveness of such a connection is pointless if the connected domains are very small, while it is amplified by domains of larger size. We capture this as follows:

Property (Size)
The size of a bisociation candidate is the number of nodes in the connected domains.

In terms of [6], the demand for size relates to the level of abstraction. A domain is more abstract than its subdomains since it includes more information units and thus an exclusive link between larger (i.e. more abstract) domains is a more promising bisociation candidate than a link between smaller domains.

Finally, the balance property assures that we avoid the situation of a very small domain attached to a large one:

Property (Balance)
A bisociation candidate is balanced iff the connected domains are of similar size.

In addition, we assume that domains of similar size tend to be of similar granularity and are thus likely to be on comparable levels of abstraction. Again, this is an approximation based on the assumption that domains are covered in comparable density. Thereby the demand for balance avoids exclusive links to small subdomains that are actually part of a broader connection between larger ones.

Following a discussion of the domain extraction process in Section 3.1, we will turn these three properties into a concrete measure for the quality of a bisociation candidate in Section 3.2.

3 Finding and Assessing Bisociations

In this section, we translate the demands described in the last section into an algorithm for the extraction and rating of bisociations. Therein we follow the

previously indicated division of tasks: (i) domain extraction and (ii) scoring of bisociation candidates.

3.1 Domain Extraction

As described in Section 2, domain affiliation of nodes is reflected by similar direct and indirect neighborhoods in the graph. Thus comparing and grouping nodes based on their neighborhoods yields domains. In the following, we establish the close relation of a node similarity measure called *activation similarity* [16] to the above described demands. Based on this similarity, we show in a second part how domains can be found using hierarchical clustering.

Activation Similarity. The employed node similarity is based on *spreading activation* processes in which initially one node is activated. The activation spreads iteratively from the activated node, along incident edges, to adjacent nodes and activates them to a certain degree as well. Given that the graph is connected, not bipartite, and activation values are normalized after each step, the process converges after sufficient iterations. The final activation states are determined by the principal eigenvector of the adjacency matrix of the underlying graph as shown in [1]. They differ, however, by their initial state and those following it. Adopting the notation of [1], activation states of all nodes at a certain time k can be represented by the activation vector $\mathbf{a}^{(k)} \in \mathbb{R}^n$ given by

$$\mathbf{a}^{(k)} = A^k \mathbf{a}^{(0)} / \left\| A^k \mathbf{a}^{(0)} \right\|_2 .$$

The value $\mathbf{a}^{(k)}$ at index v, i.e. $\mathbf{a}_v^{(k)}$, is the activation level of node v and the initial activation levels of all nodes are determined by $\mathbf{a}^{(0)}$. The denominator in the equation further ensures that the overall activation levels do not grow unrestricted with the dominating eigenvalue. We add a parameter u to denote that the spreading activation was started by activating node u with unit value. The level of activation $\mathbf{a}_v^{(k)}(u)$ of a certain node $v \in V$ at a time k, induced by a spreading activation process started at node u, reflects the reachability of node v from node u via walks of length k. More precisely, it represents the fraction of weighted walks of length k from u to v among all walks of length k started at u. The more walks end at v the better is v reachable from u and the higher its activation level $\mathbf{a}_v^{(k)}(u)$ will be. To consider more than just walks of a certain length, the activation vectors are normalized and accumulated. In this accumulation an additional decay $\alpha \in [0, 1)$ serves to decrease the impact of longer walks. The *accumulated activation vector* of node u is then defined by

$$\hat{\mathbf{a}}^*(u) = D^{-\frac{1}{2}} \left(\sum_{k=1}^{k_{\max}} \alpha^k \mathbf{a}^{(k)}(u) \right),$$

with $D = \mathrm{diag}(d(v_1), \ldots, d(v_n))$ being the degree matrix and $k_{\max}$ the number of spreading iterations. Using $D^{-\frac{1}{2}}$ for degree normalization accounts for nodes

of a very high degree: these are more likely to be reached and would thus distort similarities by attracting the activation if not taken care of. The value $\hat{\mathbf{a}}^*_v(u)$ represents the (normalized) sum of weighted walks of different lengths $1 \leqslant k \leqslant k_{\max}$ from u to v proportional to all weighted walks of different length starting at u and thus the relative reachability from u to v. $\hat{\mathbf{a}}^*(u)$ consequently serves as a description of the relations of node u to all other nodes in the graph.

Our basic assumption was, that nodes of similar domain are strongly connected and have a strong overlap of direct and indirect neighborhood. Hence, their reachability among each other is higher than that to other nodes. A comparison of the accumulated activation vectors of nodes compares the reachability of all other nodes from the specific nodes. On this basis we define the activation similarity $\sigma_{\text{act}} : V \times V \to \mathbb{R}$ between nodes $u, v \in V$ as

$$\sigma_{\text{act}}(u, v) = \cos(\hat{\mathbf{a}}^*(u), \hat{\mathbf{a}}^*(v))$$

and use it as node similarity for domain identification. For usual reasons we use the corresponding distance $1 - \sigma_{\text{act}}(u, v)$ for hierarchical clustering.

Domain Identification. Based on the distance described above, we apply hierarchical clustering for domain identification. To decide which subsets are to be merged we use Ward's linkage method [17], which minimizes the sum of squared distances within a cluster. This corresponds well with the notion of a domain since it tends to produce compact clusters and to merge clusters of similar size. First of all, we would expect a certain amount of similarity for arbitrary information units within a domain and thus a compact shape. Further, clusters of similar size are likely to represent domains on the same level of granularity and thus merging those corresponds to building upper-level domains. The resulting *merge tree* is formalized as follows:

Definition 2 (Merge tree). *A merge tree $T = (V_T, E_T)$ for a graph $G = (V, E)$ is a tree produced by a hierarchical clustering with node set $V_T = V \cup \Lambda$ where Λ is the set of clusters obtained by merging two nodes, a node and a cluster or two clusters. E_T describes the merging structure: $\{u\lambda, v\lambda\} \subseteq E_T$ iff the nodes or clusters u and v are merged into cluster $\lambda \in \Lambda$.*

However, not all clusters in the hierarchy are good domain candidates. If a cluster is merged with a single node, the result is unlikely to be an upper-level domain. Most likely, it is just an expansion of an already identified domain resulting from agglomerative clustering. These considerations lead to the domain definition:

Definition 3 (Domain). *A cluster δ_1 is a domain iff it is merged with another cluster in the corresponding merge tree:*

$$\delta_1 \in \Lambda \text{ is a domain} \Leftrightarrow \exists \delta_2, \kappa \in \Lambda \text{ such that } \{\{\delta_1, \kappa\}, \{\delta_2, \kappa\}\} \subseteq E_T \ .$$

Note that in this definition δ_2 is also a domain.

3.2 Scoring Bisociation Candidates

In the next step, we iterate over all pairs of disjoint domains and construct a bisociation candidate for each pair. We then try to assess the potential of each candidate using the properties shown in Section 2. A first step in this assessment is the identification of the bridging nodes:

Definition 4 (Bridging nodes). *Let δ_1 and δ_2 be two domains derived from the merge tree of the graph $G = (V, E)$. The set of bridging nodes* $\mathrm{bn}(\delta_1, \delta_2)$ *containes all nodes that are connected to both domains:*

$$\mathrm{bn}(\delta_1, \delta_2) = \{v \in V : \exists \{v, u_1\}, \{v, u_2\} \in E \text{ with } u_1 \in \delta_1, u_2 \in \delta_2\} .$$

Note that this definition includes nodes belonging to one of the two domains, thereby allowing direct connections between nodes of these domains.

Using this concept we can define the *b-score*, which combines the properties described in Section 2 into a single index that can be used to compare bisociation candidates directly. We therefore consider each property separately and combine them into an index at the end.

Exclusiveness could be directly expressed by the number of nodes in $\mathrm{bn}(\delta_1, \delta_2)$. However, this is not a sufficient condition. Nodes of high degree are likely to connect different domains, maybe even some of them exclusively. Nevertheless, such nodes are unlikely to form good bisociations since they are not very specific. On the other hand, bridging nodes providing only a few connections at all (and thus a large fraction of them within δ_1 and δ_2) tend to express a very specific connection. This interpretation of node degrees is of course an unproved assumption, yet we consider it as necessary and reasonable. Since we are only interested in the case of specific connection, we assess exclusiveness by using the inverse of the sum of the bridging nodes' degrees: $2/\sum_{v \in \mathrm{bn}(\delta_1, \delta_2)} d(v)$. The 2 in the numerator ensures that this quantity is bound to the interval $[0, 1]$, with 1 being the best possible value. The balance property is accounted for by relating the domain sizes in a fraction: $\min\{|\delta_1|, |\delta_2|\}/\max\{|\delta_1|, |\delta_2|\}$, again bound to $[0, 1]$ with one expressing perfect balance. Finally, the size property is integrated as the sum of the domain sizes.

As described above, a combination of all three properties is a necessary prerequisite for an interesting bisociation. Therefore, our bisociation score is a product of the individual quantities. Only in the case of $\mathrm{bn}(\delta_1, \delta_2) = \emptyset$ is our measure undefined. However, this situation is only possible if the domains are unconnected, so we define the score to be 0 in this case. For all non-trivial cases the score has strictly positive values and is defined as follows:

Definition 5 (b-score). *Let δ_1 and δ_2 be two domains, then the b-score of the corresponding bisociation candidate is*

$$b\text{-}score(\delta_1, \delta_2) = \frac{2}{\sum\limits_{v \in bn(\delta_1, \delta_2)} d(v)} \cdot \frac{\min\{|\delta_1|, |\delta_2|\}}{\max\{|\delta_1|, |\delta_2|\}} \cdot (|\delta_1| + |\delta_2|).$$

This combination acts comparably to a conjunction of the involved properties. In our opinion, an ideal bisociation is represented by two equally sized domains connected directly by a single edge or indirectly by a node connected to both domains. This optimizes the b-score, leaving the sum of the domain sizes as the only criterion for the assessment of this candidate. Further, every deviation from this ideal situation results in a deterioration of the score. In addition, the calculation of the b-score only involves information about the two domains and their neighborhoods and not the whole graph, which is important when the underlying graph is very large.

3.3 Complexity and Scalability

To determine the complexity and scalability of the complete algorithm, the process can be split into three parts: similarity computation, clustering, and scoring of domain pairs. In the following, we examine the complexity of each of these three parts.

To compute the pairwise activation similarities, the accumulated activation vectors for all nodes need to be determined. For each node several matrix vector multiplications, normalizations, scalings and additions are necessary. The computational complexity is dominated by that of the matrix vector multiplication with a complexity of $\mathcal{O}(n^2)$. Repeating this process for all nodes leads to an overall complexity of $\mathcal{O}(n^3)$.

Note, that this is a worst case result which can be improved substantially by exploiting the graph structure and the characteristics of the convergence of the spreading activation process: First of all, in a large, sparsely connected network the activation is only spread over existing edges. This alone speeds up the matrix multiplication depending on the network density which should usually be low, since otherwise the considered information units are connected to most other information units which is not very informative. Further, a large network diameter yields strongly localized activation vectors (i.e. most nodes have zero activation) in the first few iterations, since activation can only spread to nodes that are adjacent to already activated nodes in each step. This can be exploited, when only the first few activation vectors are used to approximate the activation similarity. In addition, the convergence rate of the power iteration itself is exponentially related to the ratio of the first two eigenvalues and the additional decay factor (c.f. [1]). The latter guarantees that only a few iterations of activation spreading are necessary and together with the sparsity and large diameter of the underlying network, only a small part has to be considered in the computation for an individual node. Unfortunately, the assumption of a large diameter is contradicted by many observed real-world networks and our own application example. A possible counter measure could be the removal of high-degree nodes, which should result in a larger diameter and only minimal information loss since the information provided by these nodes is most likely of highly general nature.

The crucial part of the domain identification process is the clustering of the nodes based on the computed similarity. Here, the complexity is dominated by

the ward clustering which involves $\mathcal{O}(n^2 \log^2 n)$ steps (c.f. [5]). This could be further reduced by the employment of other clustering algorithms (i.e. density based approaches) or a completely different domain identification process.

The final step is the determination of the b-scores for all domain pairs. Since in the hierarchical clustering $n-1$ merge steps are executed, $|V_T|$ and thus the number of domains is bound by $n-1$. Consequently, the number of domain pairs to be analyzed is less than $\binom{n}{2}$, i.e. in $\mathcal{O}(n^2)$. For the determination of the b-score of an individual domain pair, the domain sizes can already be prepared in the domain identification process, avoiding additional time consumption. The complexity of the b-score computation for two domains δ_1, δ_2 is therefore determined completely by the calculation of $\sum_{v \in bn(\delta_1,\delta_2)} d(v)$ - the sum of the bridging nodes' degrees. This calculation is again dominated by the determination of the elements of $bn(\delta_1, \delta_2)$. Considering a domain as a set of its contained nodes these common neighbors can be determined in $\mathcal{O}\left(\max_{\delta \in \{\delta_1,\delta_2\}} \sum_{v \in \delta} d(v)\right)$ as shown by Algorithm 1.

Algorithm 1. b-score computation

```
Input: domains δ1, δ2, graph G = (V, E)
Result: b-score(δ1, δ2)
for v ∈ δ1 do
    for u ∈ N(v) do
        m_u := true;
s := 0;
for v ∈ δ2 do
    for u ∈ N(v) do
        if m_u then
            s := s + d(u);
            m_u := false;
return 2/s · min(|δ1|,|δ2|)/max(|δ1|,|δ2|) · (|δ1| + |δ2|);
```

With $m_v = \text{false}\ \forall v \in V$ being initialized only once for the whole computation and cleaned up after each candidate evaluation, the complexity of the procedure is directly related to the loops over the neighbors of each node in either domain.

Besides the clustering process, the determination of b-scores is an important aspect in the total time spent in the analysis of a dataset. To speed up this process, we propose pruning of the set of domains and candidate domain pairs. Recall our definition of a domain based on the merge tree: it is sufficient that a cluster from Λ is merged with another cluster from Λ in contrast to merging with a single element from V. Firstly, this produces a large number of small domains: e.g. two node domains which are in turn merged with other elements from Λ. Secondly, this procedure yields a number of largely overlapping clusters that differ only in a small numer of nodes, e.g. when a large cluster is merged with a very small one. This is illustrated by the distribution of domain sizes in

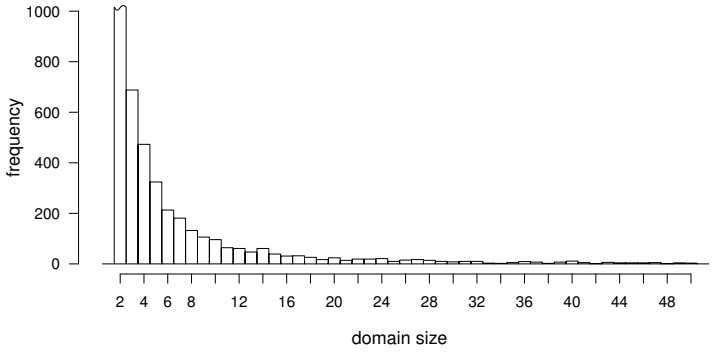


Fig. 1. Distribution of domain sizes for domains of size 0 to 50 (the number of 2-node domains is not depicted completely)

our evaluation example of Section 4 shown in Figure 1. It can be observed that a large number of domains consist of only two or three nodes. Considering the demand for size, balance and the exclusiveness of a connection between such small domains it can be seen that a large gain in efficiency could be obtained by pruning small domains or bisociation candidates involving very small domains.

In addition, a further reduction of the number of domain pairs to be considered may be achieved by filtering highly unbalanced candidates, though in that case a threshold needs to be chosen cautiously.

4 Preliminary Evaluation

To demonstrate our approach, we applied our method to the Schools-Wikipedia[1] (2008/09) dataset, which is described in more detail in [16]. Due to the lack of a benchmark mechanism we manually explored the top rated bisociation candidates and describe some of them to demonstrate the reasonability of the results.

The dataset consists of a subset of the English Wikipedia with about 5500 articles. For our experiment, we consider each article as a separate unit of information and model it as a node. We interpret cross-references as relations and introduce an undirected edge whenever one article references another. The resulting graph is connected except for two isolated nodes which we removed beforehand. For the remaining nodes we extracted the domains using the procedure described above.

Parameter Choices. To focus on the local neighborhood of nodes we used the decay value $\alpha = 0.3$ for the activation similarity. Due to this decay and the graph structure the activation processes converged quickly allowing a restriction to $k_{\max} = 10$ iterations for each process. This choice seems arbitrary, but we ensured

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

that additional iterations do not contribute significantly to the distances. First of all, the values of the following iterations tend to vanish due to the exponentially decreasing scaling factor, e.g. 0.3^{10} in the last iteration. Additionally, we ensured that the order of distances between node pairs is not altered by further iterations.

Domain Results. Altogether, we extracted 4,154 nested domains resulting in 8,578,977 bisociation candidates.

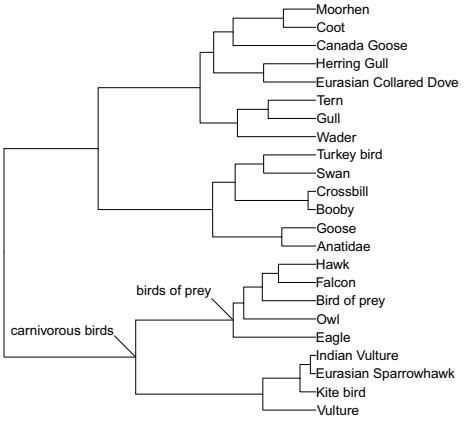


Fig. 2. Part of merge tree with articles about birds

A part of the merge tree involving birds is shown in Figure 2. This small excerpt illustrates that the hierarchical clustering can yield conceptually well defined domains, though we could not verify the complete result manually. In the example, birds of prey such as hawk, falcon, eagle etc. end up in the same cluster with carnivorous birds such as e.g. vulture and are finally combined with non-carnivorous birds to a larger cluster. This example further illustrates that the nodes of a good domain are not necessarily connected, as there are few connections within the sets of birds, and yet they share a number of external references.

Bisociation Results. The b-scores of the best 200 bisociation candidates are shown in Figure 3. It can be observed that the scores quickly decrease from some exceptionally high rated examples (b-score 1.5 and more) to the vast majority of candidates rated lower than 1. This indicates that - using the b-score as basis of judgement - the dataset contains some outstanding bisociations while most candidates are uninteresting. Since the individual candidates have to be assessed manually, this encourages the decision to concentrate on the first few pairs. Note, that due to the design of the b-score the best rated candidates often exhibit only a single bridging node. In addition, these bridging nodes appear repeatedly in bisociation candidates that differ only slightly in the composition of the connected domains which usually shrink along with decreasing b-scores.

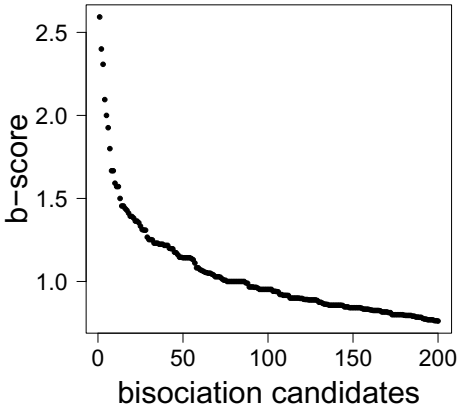


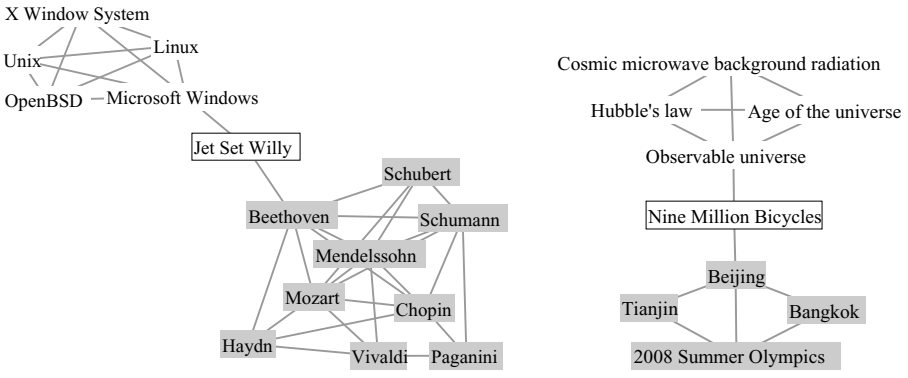
Fig. 3. Distribution of the b-score for the 200 top rated bisociation candidates

In such cases we focused on the first appearance and ignored the lower rated ones. Finally, due to the employed distance in the domain extraction process the resulting domains are not necessarily connected.

Result Evaluation. Since a comprehensive description and illustration of results would quickly exceed the scope of this article, we only show the three top-rated candidates and emulate a realistic result evaluation by additionally presenting interesting candidates found under the top-rated ones. In our visualizations of the individual candidates, we show the nodes of the individual candidate together with the link structure of the extracted network. Domain affiliation of the nodes is indicated by their background color (white or gray) while bridging nodes are highlighted by a black node border.

The overall best rated bisociation candidate, shown in Figure 4a, is composed of a domain of classical music composers such as *Schumann*, *Schubert*, *Beethoven*, *Mozart*, etc. connected to a domain incorporating operating systems and software such as *Microsoft Windows*, *Linux*, *Unix* and the *X window system*. Intuitively these domains - composers and operating systems - are highly separated and a direct connection seems to be unlikely. However, a connection between both is provided by *Jet Set Willy* which is a computer game for the *Commodore 64*.

The unusual connection to the domain of composers is explained by its title music, which was adapted from the first movement of *Beethoven*'s Moonlight Sonata. On the other side, a level editor developed for *Microsoft Windows* connects to the domain of operating systems. To us, this connection was new and surprising, though one might argue its value. Besides that, the formalized demands are met well. The connection itself is very specific, since *Jet Set Willy* provides only a few links and the two domains are far apart, i.e. not connected otherwise. The sizes of the two domains are not exactly equal but with 5 nodes



(a) b-score=2.59 (b) b-score=2.00

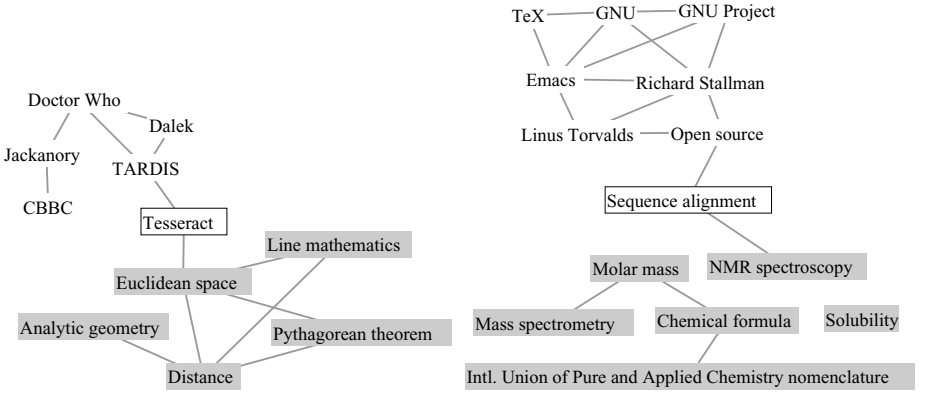
Fig. 4. The two top rated bisociations and their b-score (see text for details)

in one domain and 9 nodes in the other they are comparable. Finally, in view of the small size of the dataset and its wide distribution of topics, the absolute domain sizes are reasonable.

Figure 4b depicts the next best candidate where the node *Nine Million Bicycles* connects a geography with an astronomy domain. Excluding repetitions of *Jet Set Willy*, this is the second best rated bisociation candidate and it also appears in several variants within the best rated candidates. The bridging node *Nine Million Bicycles* refers to a song which was inspired by a visit in *Beijing*, China. The connection to the astronomy section is established by the second verse which states that the *Observable universe* is twelve billion years old. The actual link in the data set is established by the report of a discussion about the correctness of this statement in the article. As in the first example, the value of this connection is at least arguable, while the formal criteria are met well.

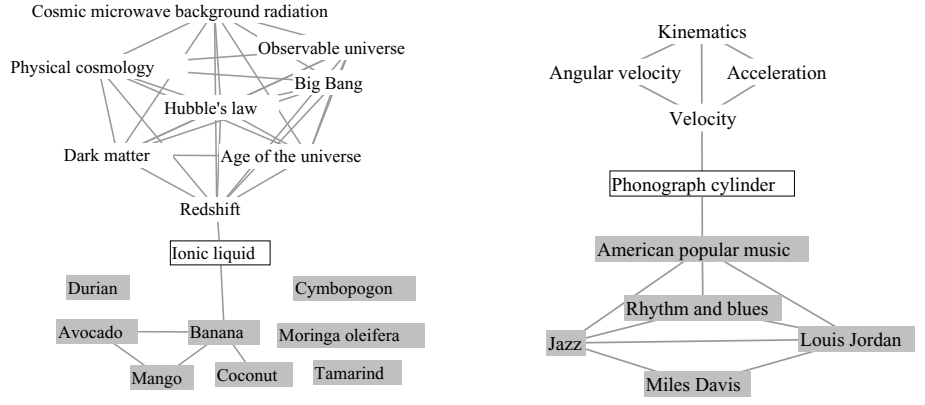
Figure 5a shows the third best rated candidate. Here, the node *Tesseract* connects a geometry domain with the domain of a famous BBC series called *Doctor Who*. A *tesseract* is a special geometrical shape also known as hypercube, which has a natural connection to the domain of geometry. In the context of the TV-series it is used to describe the form of *Doctor Who*'s spaceship called *TARDIS*.

In the following, we present some hand-picked samples that appeared as interesting to us. These should illustrate, that despite the limits imposed by the analyzed data some interesting, though not always valuable, proposals where made by the presented method. Figure 5b shows a bisociation candidate, where the node *Sequence alignment* connects domains from computer science and chemistry. The connection to the computer science domain results from reports about open source software programs implementing some of the involved algorithms. *NMR spectroscopy*, providing the connection to a domain about chemistry, is an analysis technique with applications in organic chemistry.



(a) b-score=1.67 (b) b-score=1.59

Fig. 5. Third best rated bisociation candidate and a bisociation candidate connecting open source related articles with articles about chemistry



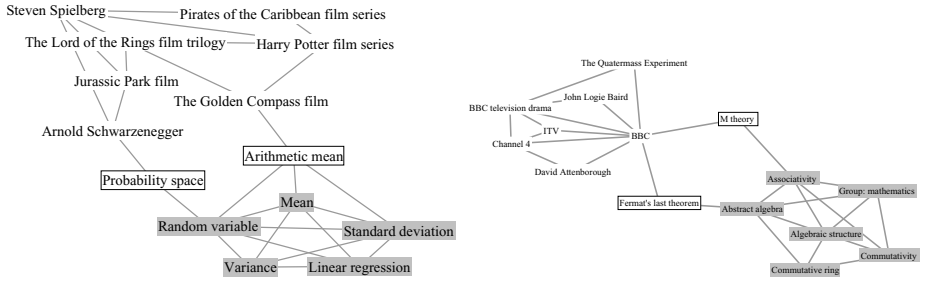
(a) b-score=1.39 (b) b-score=1.31

Fig. 6. Two interesting bisociation candidates with good b-scores

A quite surprising example - at least for us - is given in Figure 6a, where *ionic liquid* provides a nearly direct connection between the *Redshift* effect and *Banana*. While the relation to the *Redshift* effect is due to applications of ionic liquids in infra-red imaging in telescopes, the link to *Banana* is produced by an example for an application in food science

An example for a historical bisociation is shown in Figure 6b, where a music domain is connected to some physical notions in an article about the *phonograph-cylinder*. The *phonograph-cylinder* was - quoting the article - the "earliest method of recording and reproducing sound" and thus a historically new connection between physics and music. The concrete connections are established by a

discussion of the physical properties of cylinders with respect to sound recording and reproduction and the fact that this technique was of major importance in the development of recorded music, discussed in the article about *American popular music*.



(a) b-score=0.63 (b) b-score=0.53

Fig. 7. Bisociation candidates with two bridging nodes

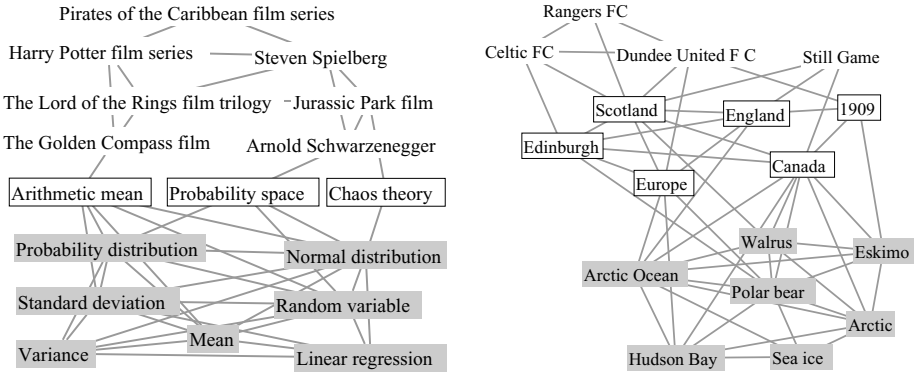
Due to the nature of the b-score, bisociation candidates with more than one bridging node tend to receive lower scores as they provide more opportunity for edges. To illustrate that there are actually interesting candidates with more than one bridging node, we show some examples for such cases in Figures 7 and 8a.

The first of these (Figure 7a) connects a domain of movies and actors with a mathematical domain. The connection itself is established by direct links, since the bridging nodes belong to the mathematical domain. Both of these links are deeply semantically reasoned, since *Schwarzenegger* is used in a voting example in the article about *Probability theory* while the link from the article *The Golden Compass film* is explanatory in the report about different ratings of the movie.

The same is true for our second example where an algebraic/mathematical domain is connected to a domain around the node *BBC*. Both connections to the *BBC* refer to documentary productions about the corresponding topics *M theory* and *Fermat's last theorem*. While the latter is clearly related to the algebraic/mathematical domain (it is a theorem of abstract algebra), the link from *M theory* to *Associativity* appears in the explanation of some details about *M theory*.

An example for a bisociation candidate with three bridging nodes and a b-score of 0.35 is shown in Figure 8a. It is basically an extension of the example shown in Figure 7a. The bridging nodes *Chaos theory*, *Probability space* and *Arithmetic mean* connect a statistics domain with nodes like *Variance* or *Mean* and a movie domain with nodes like *Arnold Schwarzenegger* or *The Lord of the Rings*.

For completeness, we additionally evaluated very low rated candidates. A negative example of a bisociation can be seen in Figure 8b. A football domain consisting of football clubs *Celtic F.C.*, *Rangers F.C.*, etc. is connected to an arctic domain containing *Arctic*, *Arctic Ocean*, *Polar bear*, etc. The bridging nodes



(a) b-score=0.35 (b) b-score=0.003

Fig. 8. A bisociation candidate (b-score=0.35) involving three bridging nodes (8a) and a bad bisociation (b-score=0.003) due to non exclusiveness and missing balance (8b)

are countries such as *Canada*, *Europe*, *England*, etc. and *Edinburgh*. Clearly, this bisociation candidate is not exclusive since the number of connecting nodes is high (proportional to the domain sizes) and the degree of these nodes is high as well (countries have a very high degree in Schools-Wikipedia).

The above examples illustrate that our index discriminates well with respect to exclusiveness and balance. A detailed examination showed in addition that size is negatively correlated with both other index components. This and the limited size of the dataset could explain the small sizes of the best rated candidates.

Our preliminary evaluation indicates the potential of the presented method to detect bisociations based on the analysis of the graph structure. Even though Schools-Wikipedia is a reasonable dataset for evaluation purposes, one cannot expect to find valuable or even truly surprising bisociations therein since it is limited to handpicked, carefully administrated common knowledge, suitable for children. We opted to manually evaluate the results since the value of a bisociation is a highly subjective semantic property, which inhibits automatic evaluation. An evaluation using synthetic data is complicated by the difficulty of realistic simulation and could introduce an unwanted bias on certain types of networks, distorting the results. Finally, manually tagged datasets for this purpose are not available.

5 Related Work

Although a wealth of techniques solving different graph mining problems already exist (see e.g. [4] for an overview), we found none to be suitable for the problem addressed here. Most of them focus on finding frequent subgraphs, which is not

of concern here. Closely related to our problem are clustering and the identification of dense substructures, since they identify structurally described parts of the graph. Yet bisociations are more complicated structures due to a different motivation and therefore require a different approach to be detected.

The exclusiveness of a connection between different groups is also of concern in the analysis of social networks. Social networks are used to model contacts, acquaintances and other relations between persons, companies or states. Burt [3] for example regards the connections in a network of business contacts as part of the capital a player brings to the competitive arena. In his setting, a player profits if he can provide an exclusive connection between two otherwise separated groups. By controlling this connection, he controls the flow of information or value between the groups, thereby gaining an advantage. Burt terms such a situation a *structural hole* that is bridged by this player. Translating the two separated groups into domains and the player into a bridging node relates Burt's concept to the bisociation. However, in the index he defines to measure the presence of a structural hole only the very local view of the player himself is integrated. Further, his index would implicitly render domains a product of only direct connections between concepts, whereas we showed earlier that a more general concept of similarity is advisable.

A global measure for the amount of control over connections between other players is provided by *betweenness* [7]. Betweenness measures the fraction of shortest paths between all other nodes that employ an individual node. Intuitively, the shortest paths in a network are the preferred channel of transportation. Consequently, if a node appears in a large fraction of these shortest connections, it can exert a certain amount of control on the flow of goods or information. The translation to our setting is again straightforward, but it provides no explanation of what a domain is. However, this approach leads to the variant of hierarchical graph clustering proposed in [12]. Girvan and Newman develop a top-down approach in which the edges of highest betweenness are removed recursively until the graph splits up into several components. Still, only a subset of the properties we demand from a bisociation is considered.

Strongly related to bisociations is the notion of serendipity [15] which describes accidental discoveries. Serendipitous discoveries strongly overlap with bisociations since the involved fortuitousness is often caused by the connection of dissimilar domains of knowledge. A number of approaches (e.g. [13,8]) were developed to implement this concept in recommender systems balancing between the suggestion of strongly related versus loosely related surprising suggestions of content which lead the user into new directions not too far from his original interests. In a sense this work is parallel to ours, but targets a different setting - users and their preferences - and thus follows different criteria of optimality.

None of these approaches provide a coherent, formal setting for the description of domains and potentially interesting links between these. Note further, that our approach is additionally distinguished from all of the mentioned variants in that the process of community or domain detection is guided by a node similarity tailored to the identification of domains of knowledge.

6 Conclusion

We presented an approach for the discovery of potentially interesting, domain crossing associations, so-called bisociations. For this purpose we developed a formal framework to describe potentially interesting bisociations and corresponding methods to identify domains and rank bisociations according to interestingness. Our evaluation on a well-understood benchmark data set has shown promising first results. We expect that the ability to point the user to potentially interesting, truly novel insights in data collections will play an increasingly important role in modern data analysis.

The presented method is, however, not intended to be directly applicable to real world problems. Instead, we presented a framework that can be used as a guideline and benchmark for further developments in this direction. Conceptually, we divide the presented approach in two parts: (i) basic considerations about the expression of domains and bisociations in a structural knowledge description and (ii) a framework for the identification of the described concepts. To demonstrate the soundness of our considerations and their general applicability, we then filled this framework with a number of heuristics to solve the resulting subproblems. Clearly, the choice of these heuristics is to some extend arbitrary and can be improved, especially in light of additional experience with more realistic data. However, by using them in a first instantiation of the described framework, we demonstrated that the underlying assumptions lead to promising results. Finally, we hope that further improvements of this framework will ultimately lead to systems that are applicable in practical settings.

Acknowledgements. This research was supported by the DFG under grant GRK 1042 (Research Training Group „Explorative Analysis and Visualization of Large Information Spaces“) and the European Commission in the 7th Framework Programme (FP7-ICT-2007-C FET-Open, contract no. BISON-211898).

References

1. Berthold, M.R., Brandes, U., Kötter, T., Mader, M., Nagel, U., Thiel, K.: Pure spreading activation is pointless. In: Proceedings of the CIKM the 18th Conference on Information and Knowledge Management, pp. 1915–1919 (2009)
2. Boden, M.A.: Précis of the creative mind: Myths and mechanisms. Behavioral and Brain Sciences 17(03), 519–531 (1994)
3. Burt, R.S.: Structural holes: the social structure of competition. Harvard University Press (1992)
4. Cook, D.J., Holder, L.B.: Mining graph data. Wiley-Interscience (2007)

5. Eppstein, D.: Fast hierarchical clustering and other applications of dynamic closest pairs. In: Proceedings of the Ninth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 1998, pp. 619–628. Society for Industrial and Applied Mathematics, Philadelphia (1998)
6. Ford, N.: Information retrieval and creativity: Towards support for the original thinker. Journal of Documentation 55(5), 528–542 (1999)
7. Freeman, L.C.: A set of measures of centrality based upon betweenness. Sociometry 40, 35–41 (1977)
8. Kamahara, J., Asakawa, T., Shimojo, S., Miyahara, H.: A community-based recommendation system to reveal unexpected interests. In: Proceedings of the 11th International Multimedia Modelling Conference (MMM 2005), pp. 433–438. IEEE (2005)
9. Koestler, A.: The Act of Creation. Macmillan (1964)
10. Kötter, T., Thiel, K., Berthold, M.R.: Domain bridging associations support creativity. In: Proceedings of the International Conference on Computational Creativity, Lisbon, pp. 200–204 (2010)
11. Maxwell, J.C.: A treatise on electricity and magnetism. Nature 7, 478–480 (1873)
12. Newman, M.E.J., Girvan, M.: Finding and evaluating community structure in networks. Physical Review E 69(2), 026113 (2004)
13. Onuma, K., Tong, H., Faloutsos, C.: Tangent: a novel, 'surprise me', recommendation algorithm. In: Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD 2009, pp. 657–666. ACM, New York (2009)
14. Poincaré, H.: Mathematical creation. Resonance 5(2), 85–94 (2000); reprinted from Science et méthode (1908)
15. Roberts, R.M.: Serendipity: Accidental Discoveries in Science. Wiley-VCH (1989)
16. Thiel, K., Berthold, M.R.: Node similarities from spreading activation. In: Proceedings of the IEEE International Conference on Data Mining, pp. 1085–1090 (2010)
17. Ward Jr., J.H.: Hierarchical grouping to optimize an objective function. Journal of the American Statistical Association 58(301), 236–244 (1963)