First publ. in: PAKDD (2008), pp. 14-25

Supporting Creativity: Towards Associative Discovery of New Insights

Michael R. Berthold, Fabian Dill, Tobias Kötter, and Kilian Thiel

University of Konstanz, Fach M712, 78484 Konstanz, Germany
Michael.Berthold@uni-Konstanz.de

Abstract. In this paper we outline an approach for network-based information access and exploration. In contrast to existing methods, the presented framework allows for the integration of both semantically meaningful information as well as loosely coupled information fragments from heterogeneous information repositories. The resulting Bisociative Information Networks (BisoNets) together with explorative navigation methods facilitate the discovery of links across diverse domains. In addition to such "chains of evidence", they enable the user to go back to the original information repository and investigate the origin of each link, ultimately resulting in the discovery of previously unknown connections between information entities of different domains, subsequently triggering new insights and supporting creative discoveries.

Keywords: BisoNet, Bisociative Information Networks, Bisociation, Discovery Support Systems.

1 Motivation: The Need for Information Exploration

Data collection and generation methods continue to increase their ability to fill up information repositories at an alarming rate. In many industries it is nowadays commonly accepted – although often not openly admitted – that only a fraction of available information is taken into account when making decisions or trying to uncover interesting, potentially crucial links between previously unconnected pieces of information.

In order to allow users to be able to find important pieces of information it is necessary to replace classical question answering systems with tools that allow for the interactive exploration of potentially related information – which can often trigger new insights and spark new ideas which the user did not expect at start and was therefore unable to formulate as a query initially. It is especially crucial for such systems to enable the seamless crossing of repository boundaries to trigger new discoveries across domains. Since we will not know at the start which types of information are needed or which kind of questions will be asked throughout this explorative process, the system always needs to be able to provide access to heterogeneous information repositories. These can be structured, well annotated repositories, such as an ontology or a database of

human annotations ("known facts") but it needs to incorporate other types of information as well, such as experimental data or the vast amounts of results from the mining of e.g. published texts ("pieces of evidence"). The real challenge lies in providing the user with easy access to all of this information so that she can quickly discard uninteresting paths to information that is not currently relevant and at the same time focus on areas of interest. Similar to drill down operations in Visual Data Mining, such a system will need to be able to show summarizations according to different dimensions or levels of detail and allow parallel changes of focus to enable the user to ultimately navigate to the information entities that explain the connections of interest. Of course, the system cannot be static but will require not only means for continuous updating of the underlying information repositories to accommodate new data, but also new and better methods to extract connections. In [1] we have argued that such a system will truly support the discovery of new insights. Related work investigating the nature of creativity (see [2] among others) describes similar requirements for creative discoveries, based on broad but at the same time context dependent more focused exploration of associations as the underlying backbone.

In this paper we outline an approach to realize such a system using a network-based model to continuously integrate and update heterogeneous information repositories and at the same time allow for explorative access to navigate both semantic and evidential links. Before describing our prototypical system in more detail we review existing network-based systems for knowledge or information modeling. We conclude the paper by discussing open issues and challenges.

2 State of the Art: Network-Based Information Access

Different network-based models have been applied to Information Retrieval, such as artificial neural networks, probabilistic inference networks, Hopfield or knowledge networks [3]. The first two are mainly used to match documents to queries and to find relevant documents related to a certain query. Documents and index terms, which are the most discriminative terms, are represented as vertices in these networks. Edges can be created to connect documents citing each other, documents with their index terms, as well as cooccurring index terms. Hopfield and knowledge networks are additionally used for automatic thesaurus creation and consultation [4]. In this case only vertices of index terms cooccurring in documents or sentences are connected via edges. Another connectionist approach, Adaptive Information Retrieval (AIR), creates additional vertices for each document author and connects them by their author co-author relationships [5,6].

The majority of these approaches use weighted networks. In these networks a weight is assigned to each edge, which depends on the underlying network model as well as the computation and interpretation of the relation. In probabilistic inference networks the weights represent probabilities of terms occurring in documents being relevant to a certain query [3,7]. Whereas the weights of knowledge or Hopfield networks as discussed in [4] represent the relatedness of cooccurring terms. Usually the weights of these approaches are only computed once and not

changed afterwards. In contrast to these approaches, Belew enables each user of an AIR model to adapt the weights according to their relevance feedback [5]. After initialization of the weights where the edges between documents and terms are weighted with the term's inverse document frequency, a user can send queries to the network. The user then rates the resulting nodes, representing terms, documents or authors, as relevant or irrelevant. This relevance feedback is passed to the network again in order to adjust the edge weight and process another query. This kind of iterative process is continued until the result fits the users needs. One essential disadvantage of such an adaptive system is that it adapts to the user's opinion of which documents are more relevant than others related to a certain query. This means that the network will, over time, be strongly biased by the opinion of the majority of the users.

In a number of other domains, networks have been applied to combine, represent, integrate and analyze information, such as bioinformatics and life science, with a strong emphasis on the extraction of pharmacological targets [8], protein functions [9], gene-gene [10], gene-protein [11] or protein-protein interactions [12,13] from different biological databases and biomedical literature [14]. To mine texts and find this kind of interaction Blaschke et al. [12] proposed to parse the sentences into grammatical units. Patterns or regular expressions have been used as well to extract genes, proteins and their relations in texts [10,13].

Once the units of information and their relations are found, they can be represented in a network. Additional algorithms can be used to cluster and analyze these networks in order to identify meaningful subnetworks (communities) [15,13]. The analysis of network structures also reveals new insights into complex processes such as regulator strategies in yeast cells [16]. Additionally the edges can be evaluated and their quality can be specified based on several features like edge reliability, relevance and rarity [17]. Note that also the increasingly popular social networks fall into this category. In general much work has been done when it comes to methods for network analysis [18].

2.1 Adaptive and Explorative Approaches

To visually analyze graphs, different layout algorithms such as the force-directed Fruchterman-Reingold algorithm [19] have been developed. But large networks with several million vertices and many more edges cannot be visualized completely in a reasonable manner. Therefore the visualization has to be focused on a subgraph or at least summarized to match the current user's interest or give an overview. Various visualization techniques have been developed to address this problem. Examples are the generalized Fisheye views [20], the splitting of a network into several smaller semantical distinct regions [21] or the interactive navigation through different levels of abstractions [22].

Another way to analyze large networks is to extract subgraphs that contain most of the relevant information. One way to do this is to query a graph. On the one hand queries can be generated by manually drawing a sub-graph or by using a particular query language, i.e. GenoLink [23]. The results of such queries are represented as sub-graphs which themselves could be the starting point of further

analyses. On the other hand Spreading Activation techniques are very common techniques to explore networks and handle queries [24]. In general the idea of activity spreading is based on assumed mechanisms of human cognitive memory operations, originated from psychological studies [25]. These techniques are adopted to many different areas such as Cognitive Science, Databases, Artificial Intelligence, Psychology, Biology and Information Retrieval. The basic activity spreading technique is quite simple. First, one or more vertices, representing the query terms, are activated. The initial activation is distributed (spread) over the outgoing edges and activates in subsequent iterations the adjacent vertices. This iterative process will continue until a certain termination condition, such as a maximum number of activated nodes or iterations or a minimum edge or vertex weight is reached. The activation itself can also be weighted and can decrease over time or when propagating over certain edges. Furthermore different activation functions can be used for the vertices [24]. In [4] the networks are explored by usage of a branch-and-bound search and a Hopfield net activation. Due to the restriction that a Hopfield activation algorithm only guarantees to converge if the graph's adjacency matrix is symmetric, meaning that the graph is undirected, this technique is only applicable for certain kinds of networks. Other approaches cope with the complexity by clustering or pruning the graph based on their topology [26] or based on additional information such as a given ontology [27].

2.2 Combining Heterogeneous Information Repositories

The integration of heterogeneous data sources facilitates insights across different domains. Such insights are important especially in complex application areas such as life sciences, which deal with different kinds of data, e.g. gene expression experiments, gene ontologies, scientific literature, expert notes, etc. During the last few years several approaches have been developed that attempt to tackle this problem. The authors of [28] classified these systems into three general classes: navigational integration, mediator-based integration and warehouse integration.

Navigational integration approaches like SRS [29], Entrez [30] and LinkDB [20] aim to integrate heterogeneous data by providing links between units of information derived from different sources. Links can be created based on database entries as well as on the similarity of the units of information, or manually by experts [20]. Most of the applications consist of one or more indexed flat files containing the relations between the different concepts.

The second category is the mediator-based integration systems such as DiscoveryLink [31], BioMediator [32], Kleisli [33] and its derivatives like TAMBIS [34] or K2 [35]. These systems act as a mediator, which maps the schema of different data sources onto a unified schema. Each query is converted and split up into a set of sub-queries, which are than redirected to the wrapper of the integrated data source. Finally the results of the sub-queries are combined to a single result and returned by the mediator.

Warehouse approaches like GUS [35], Atlas [36], BIOZON [37] and BNDB [38] are similar to the mediator-based approach since they also provide a unified

schema for all data sources. But instead of creating a sub-query for each data source the data itself is loaded into the unified schema.

Navigational integration and mediator-based approaches do not integrate all the detailed data of a concept. The amount and complexity to handle additional data is much smaller in comparison to systems that integrate the detailed information of a concept like the warehouse approach. The advantage of this kind of light integration is the ability to keep the detailed information up to date since it is stored in the external sources itself. The drawback of such an integration is the dependency on all the integrated systems with respect to reliability and performance. In contrast, the warehouse approach also integrates all the detailed information from the distributed repositories. The data can be preprocessed and enriched with additional information such as similarity measures or user annotations. The local storage of all data leads to a better performance and system reliability. However the huge amount of data itself and continued maintenance to detect changes and inconsitencies are the major drawback of such systems.

In summary, warehouse and mediator-based approaches provide the user with a unified, mostly relational schema. This allows professional users the ability to use powerful query languages like SQL to perform complex joins and queries. The unification leads mostly to a complex data model including link tables to combine the different data sources. Navigational approaches only maintain link information between concepts and provide simple point and click interfaces visualizing links between them. These interfaces are also manageable by semi professional users but restricted in their query capabilities like the lack of complex joins. A common goal of all the mentioned integration approaches is the combination of equal or similar concepts from different data sources. An obvious approach to link these concepts is the usage of a flexible graph structure. An example of integrating high confidence biological data is PathSys [39]. PathSys is a graph-based data warehouse, which is used to analyze relations between genes and proteins. To predict protein-protein interactions several approaches adopted Bayesian Networks to model the mostly noisy or uncorrelated evidences of biological experiments [40,41].

3 BisoNets: Bisociative Information Networks

As we have suggested above, simply finding classical associations is not sufficient to detect interesting connections across different information repositories and contexts. Existing systems either tend to be to application focussed or restricted to only a few type of information sources or types. However, in order to support creative discoveries across domains we cannot assume that we know from the beginning which information repositories will need to be combined in which way.

In 1964 Arthur Koestler introduced the term *bisociation* [42] to indicate the *"...joining of unrelated, often conflicting information in a new way..."*. Using this terminology we use the term Bisociative Information Networks, or short *BisoNets* to denote a type of information network addressing the above concerns, fusing the following requirements:

- Heterogeneous Information: BisoNets integrate information from various information repositories, representing both semantically solid knowledge (such as from an ontology or a human annotated semantic net) and imprecise and/or unreliable knowledge such as derived from automatic analysis methods (e.g. results from text mining or association rule analyses) or other experimental results (e.g. correlations derived from protein expression experiments).
- Merging Evidence and Facts: BisoNets provide a unifying mechanism to combine these different types of information and assign and maintain edge weights and annotations in order to allow the mixing of links with different degrees of certainty.
- Continuous Update: BisoNets can be refined online and continuously integrate updated or new information.
- Exploration/Navigation: Finally, in order to allow access to the resulting information structure, BisoNets provide explorative navigation methods, which show summarizations of (sub-) networks, and allow the changing of focus and quick zooming operations.

There is strong evidence that such a complex system of loosely, not necessarily semantically coupled information granules exhibits surprisingly sophisticated features. In [43] Hecht-Nielsen describes a network which generates grammatically correct and semantically meaningful sentences purely based on links created from word co-occurrence without any additional syntactical or semantical analysis. In addition, [2] discusses requirements for creativity, supporting this type of domain bridging bisociations.

3.1 First Steps: A BisoNet Prototype

In order to evaluate the concept of BisoNets, we have implemented a first prototype and so far have mainly applied it to life science related data. However, the toolkit is not restricted to this type of data. The BisoNet prototype creates one vertex for each arbitrary unit of information, i.e. a gene or protein name, a specific molecule, an index term or a document, and other types of named entities. Relations between vertices are represented by edges. Vertices are identified by their unique name and edges by the vertices they connect. In order to model not only facts but also more or less precise pieces of evidence, edges are weighted to reflect the degree of certainty and specificity of the relation.

Due to the uniqueness of a vertex name, a vertex can be ambiguous and represent different units of information, i.e. a vertex can represent a term extracted from a document and a gene or protein name derived from a certain database. For example a vertex could represent the animal “jaguar” or the make of car. To distinguish the different kinds of meanings, an annotation can be applied to vertices and edges. An annotation specifies the origin and the type of the information unit. A vertex representing different units of information will contain different annotations: one annotation for each meaning. Edges with different annotations represent relations derived from different data sources. Each

annotation of an edge contains its own weight in order to specify the evidence of the relation according to the data sources it was derived from.

The structure of the knowledge network is rather lightweight, that is it simply consists of vertices and edges, but contains no detailed information of the vertices or edges itself. In order to access this valuable, more detailed information as well, so-called data agents have been implemented. For each annotation, representing a particular kind of information of a certain data source, a data agent is available, which can be used to access the corresponding data source and extract the detailed information for a particular vertex or edge annotation.

To analyze and explore the network in order to find new and hopefully useful information, potentially uninteresting information has to be filtered. The prototype provides several filtering methods. One method allows particular annotation types of vertices and edges to be hidden, such as terms, species, or chemical compounds to focus on a specific context. Another one filters edges by their weight to filter out all relations below a certain degree of evidence. To extract information related to a particular issue, an activity spreading algorithm has been implemented, similar to the branch-and-bound algorithm of [4], which is able to extract subgraphs consisting of the most relevant vertices related to a specified set of initially activated vertices.

We implemented the BisoNet prototype within the modular information mining platform KNIME [44] due to the large set of data preprocessing and analysis methods available already. Each procedure and algorithm dealing with the network was implemented as a module or KNIME node respectively. This allows them to be used and combined individually and networks can be created, analyzed and explored in a flexible manner. Figure 1 shows an example KNIME workflow in which a network was created consisting of PubMed [45] abstracts as text data, gene subgroup information derived from gene expression data, gene-gene interaction data from Genetwork [46] and Gene Ontology [47] information. One by one all data sources are integrated into the network and at the end of the pipeline various filters can be applied to concentrate on a particular subgraph.

To visualize the network we used Cytoscape [48] an open source software platform for graph visualization. Note that this graph visualization toolkit does not offer sophisticated means to navigate the underlying BisoNet.

To create the complete network PubMed abstracts, related to the drug Plavix, treating thrombotic events, were analyzed and all content bearing index terms, gene and compound names were extracted and inserted into the network as vertices. Co-occurring terms above a certain frequency are connected by an edge. In addition gene-gene interaction data of Genetwork was integrated and, by applying different filters such as gene annotation filter or edge weight filter, the subgraph shown in Figure 2 can be extracted. The graph consists of 27 vertices representing gene names and 33 edges representing gene-gene interactions. The green vertices stem from the Genetwork data, the brown vertices from PubMed text data. In the subgraph illustrated in Figure 2 the four genes derived from text data connect and supplement the gene subgraphs of the Genetwork data nicely. Note how connections between subgraphs based on one data source are connected by information derived from a second source.

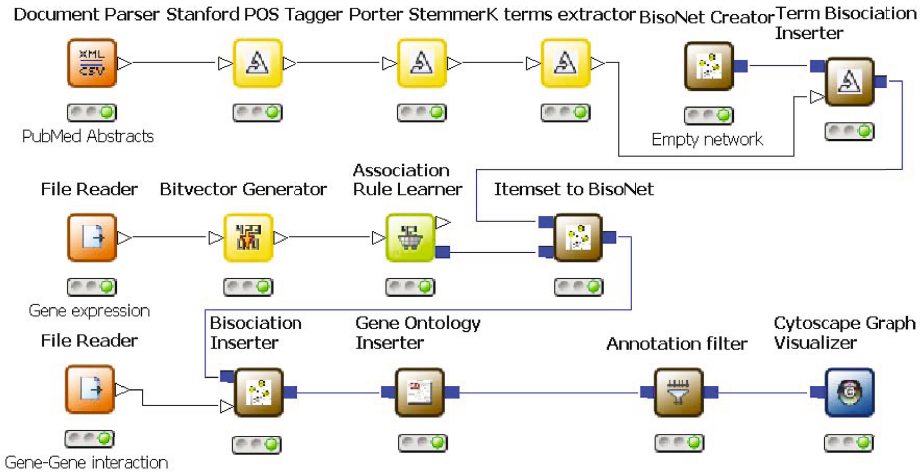

Fig. 1. A KNIME workflow which creates a network consisting of text and gene data. See text for details.

3.2 Open Issues and Challenges

The BisoNet prototype as described above is a first attempt at implementing the concepts listed in Section 1. Many open issues and challenges are still awaiting solutions and usable realizations. Within the EU Project "BISON" many of these challenges will be tackled over the coming years, focussing among others on issues related to:

- Scalability: addressing problems related to the increasing size of the resulting networks demanding new approaches for the storage, access, and subgraph operations on distributed representations of very large networks,
- Weight and Network Aggregation: that is, issues related to information sources of vastly different context and levels of certainty but also presumably simple problems of different versions of the same information repository, which also requires dealing with outdated information.
- Graph Abstraction: relating to methods that are especially crucial for problems related to exploration and navigation. In order to support zoom in and out operations, we need sophisticated methods for graph summarization and abstraction allowing for the offering, creation, and formalization of different views along different dimensions and at different levels of granularity on (sub) graphs.
- Disambiguation: that is, the differentiation of named entities with different meaning will also be critical to avoid nonsensical paths. Some of this will manifest automatically by supporting links of different domains but some means of at least semi automatic detection of ambiguous terms will be needed.

Without doubt, many other issues will be encountered along the way and soon cognitive issues will also become increasingly important, i.e., developing interfaces

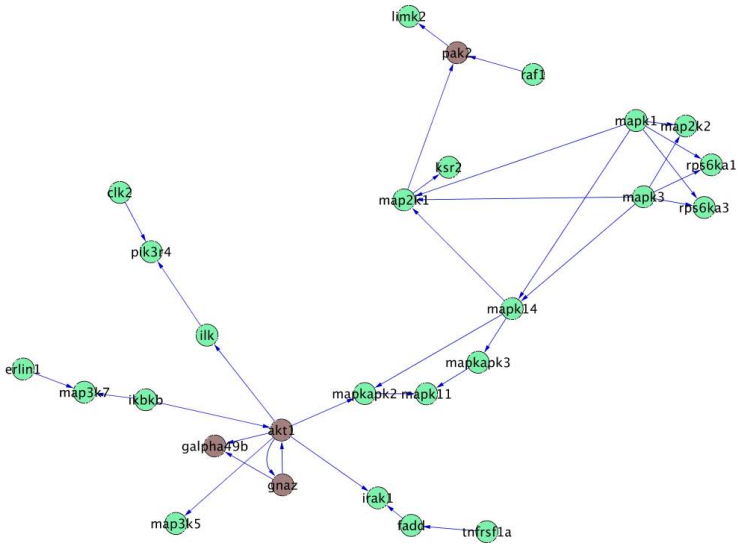

Fig. 2. A gene subgraph extracted from a network. See text for details.

that are adopted to the way humans think and work and therefore truly support human creativity instead of asking the user to adopt to the way the system has been designed.

4 Summary

In this paper we have outlined a new approach to support associative information access, enabling the user to find links across different information repositories and contexts. The underlying network combines pieces of information of various degrees of precision and reliability and allows for the exploration of both connections and original information fragments. We believe these types of bisociative information networks are a promising basis for the interactive exploration of loosely connected, semi- or unstructured information repositories, ultimately leading to fully fledged Discovery Support Systems.

Acknowledgements. We would like to thank the members of the European Framework 7 project BISON for many stimulating discussions, which have helped to refine the concept of BisoNets.

References

1. Berthold, M.R., Nürnberger, A.: Towards associative information access. In: Proceedings of AISB 2006 (Adaptation in Artificial and Biological Systems), Society for the Study of Artificial Intelligence and the Simulation of Behaviour, University of Bristol, UK, vol. 3, pp. 98–101 (2006)

2. Sternberg, R.J. (ed.): Handbook of Creativity. Cambridge University Press, Cambridge (1999)
3. Cunningham, S., Holmes, G., Littin, J., Beale, R., Witten, I.: Applying connectionist models to information retrieval. In: Amari, S., Kasobov, N. (eds.) Brain-Like Computing and Intelligent Information Systems, pp. 435–457. Springer, Heidelberg (1997)
4. 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. J. Am. Soc. Inf. Sci. 46, 348–369 (1995)
5. Belew, R.K.: Adaptive information retrieval: using a connectionist representation to retrieve and learn about documents. In: SIGIR 1989: Proceedings of the 12th annual international ACM SIGIR conference on Research and development in information retrieval, pp. 11–20. ACM Press, New York (1989)
6. Belew, R.K. (ed.): Finding Out About: A Cognitive Perspective on Search Engine Technology and the WWW, 1st edn. Cambrigde University Press, Cambrigde (2000)
7. Fuhr, N.: Probabilistic models in information retrieval. The Computer Journal 35(3), 243–255 (1992)
8. Paolini, G., Shapland, R., van Hoorn, W., Mason, J., Hopkins, A.: Global mapping of pharmacological space. Nature Biotechnology 24, 805–815 (2006)
9. Sharan, R., Ulitsky, I., Shamir, R.: Network-based prediction of protein function. Mol Syst Biol 3, 88 (2007)
10. Natarajan, J., Berrar, D., Dubitzky, W., Hack, C., Zhang, Y., DeSesa, C., Brocklyn, J.R.V., Bremer, E.G.: Text mining of full-text journal articles combined with gene expression analysis reveals a relationship between sphingosine-1-phosphate and invasiveness of a glioblastoma cell line. BMC Bioinformatics 7, 373 (2006)
11. Chiang, J.H., Yu, H.C.: Meke: discovering the functions of gene products from biomedical literature via sentence alignment. Bioinformatics 19(11), 1417–1422 (2003)
12. Blaschke, C., Andrade, M.A., Ouzounis, C., Valencia, A.: Automatic extraction of biological information from scientific text: protein-protein interactions. In: Proc. Int. Conf. Intell. Syst. Mol. Biol., pp. 60–67 (1999)
13. Hu, X., Wu, D.D.: Data mining and predictive modeling of biomolecular network from biomedical literature databases. IEEE/ACM Trans Comput Biol Bioinform 4(2), 251–263 (2007)
14. Roberts, P.M.: Mining literature for systems biology. Brief Bioinform 7(4), 399–406 (2006)
15. Newman, M.E.J.: The structure and function of complex networks. SIAM Review 45, 167 (2003)
16. Lee, T.I., Rinaldi, N.J., Robert, F., Odom, D.T., Bar-Joseph, Z., Gerber, G.K., Hannett, N.M., Harbison, C.T., Thompson, C.M., Simon, I., Zeitlinger, J., Jennings, E.G., Murray, H.L., Gordon, D.B., Ren, B., Wyrick, J.J., Tagne, J.B., Volkert, T.L., Fraenkel, E., Gifford, D.K., Young, R.A.: Transcriptional regulatory networks in saccharomyces cerevisiae. Science 298, 799–804 (2002)
17. 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)
18. Brandes, U., Erlebach, T.: Network Analysis: Methodological Foundations. Springer, Heidelberg (2005)

19. Fruchterman, T.M.J., Reingold, E.M.: Graph drawing by force-directed placement. Software-Practice And Experience 21, 1129–1164 (1991)
20. Fujibuchi, W., Goto, S., Migimatsu, H., Uchiyama, I., Ogiwara, A., Akiyama, Y., Kanehisa, M.: Dbget/linkdb: an integrated database retrieval system. Pac. Symp. Biocomput., 683–694 (1998)
21. Shneiderman, B., Aris, A.: Network visualization by semantic substrates. IEEE Trans Vis Comput Graph 12(5), 733–740 (2006)
22. Abello, J., Abello, J., Korn, J.: Mgv: a system for visualizing massive multidigraphs. Transactions on Visualization and Computer Graphics 8(1), 21–38 (2002)
23. 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)
24. Crestani, F.: Application of spreading activation techniques in informationretrieval. Artif. Intell. Rev. 11(6), 453–482 (1997)
25. Rumelhart, D., McClelland, J.: Parallel distributed processing: explorations in the microstructure of cognition, foundations, vol. 1. MIT Press, Cambridge (1986)
26. van Ham, F., 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)
27. Shen, Z., Ma, K.L., Eliassi-Rad, T.: Visual analysis of large heterogeneous social networks by semantic and structural abstraction. IEEE Trans Vis Comput Graph 12(6), 1427–1439 (2006)
28. Hernandez, T., Kambhampati, S.: Integration of biological sources: current systems and challenges ahead. SIGMOD Rec 33(3), 51–60 (2004)
29. Etzold, T., Argos, P.: Srs–an indexing and retrieval tool for flat file data libraries. Comput Appl Biosci 9(1), 49–57 (1993)
30. Schuler, G., Epstein, J., Ohkawa, H., Kans, J.: Entrez: molecular biology database and retrieval system. Methods Enzymol 266, 141–162 (1996)
31. Haas, L., Schwarz, P., Kodali, P., Kotlar, E., Rice, J., Swope, W.: Discoverylink: A system for integrated access to life sciences data sources. IBM Systems Journal 40(2), 489–511 (2001)
32. Wang, K., Tarczy-Hornoch, P., Shaker, R., Mork, P., Brinkley, J.F.: Biomediator data integration: beyond genomics to neuroscience data. In: AMIA Annu Symp Proc, pp. 779–783 (2005)
33. Chung, S.Y., Wong, L.: Kleisli: a new tool for data integration in biology. Trends Biotechnol 17(9), 351–355 (1999)
34. Stevens, R., Baker, P., Bechhofer, S., Ng, G., Jacoby, A., Paton, N.W., Goble, C.A., Brass, A.: Tambis: transparent access to multiple bioinformatics information sources. Bioinformatics 16(2), 184–185 (2000)
35. Davidson, S., Crabtree, J., Brunk, B., Schug, J., Tannen, V., Overton, G., Stoeckert, C.: K2/kleisli and gus: Experiments in integrated access to genomic data sources. IBM Systems Journal 40(2), 512–531 (2001)
36. Shah, S.P., Huang, Y., Xu, T., Yuen, M.M.S., Ling, J., Ouellette, B.F.F.: Atlas - a data warehouse for integrative bioinformatics. BMC Bioinformatics 6, 34 (2005)
37. Birkland, A., Yona, G.: Biozon: a system for unification, management and analysis of heterogeneous biological data. BMC Bioinformatics 7, 70 (2006)
38. Kuentzer, J., Backes, C., Blum, T., Gerasch, A., Kaufmann, M., Kohlbacher, O., Lenhof, H.P.: Bndb - the biochemical network database. BMC Bioinformatics 8, 367 (2007)

39. 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)
40. 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(5644), 449–453 (2003)
41. 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)
42. Koestler, A.: The Act of Creation. London Hutchinson (1964)
43. Hecht-Nielsen, R.: 3. In: Confabulation Theory, pp. 73–90. Springer, Heidelberg (2007)
44. Berthold, M.R., Cebron, N., Dill, F., Gabriel, T.R., Kötter, T., Meinl, T., Ohl, P., Sieb, C., Thiel, K., Wiswedel, B.: KNIME: The Konstanz Information Miner. In: Studies in Classification, Data Analysis, and Knowledge Organization (GfKL 2007), Springer, Heidelberg (to appear)
45. Of Medicine, U.N.L.: Pubmed (last accessed January 11, 2008), `http://www.ncbi.nlm.nih.gov/sites/entrez/`
46. 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. Am J Hum Genet 78, 1011–1025 (2006)
47. Consortium, G.O.: Creating the gene ontology resource: design and implementation. Genome Res 11(8), 1425–1433 (2001)
48. Cytoscape: Cytoscape (last accessed January 11, 2008), `http://www.cytoscape.org/`