Abstract of the Dissertation

Visual Data Mining of Graph-Based Data

Oliver Niggemann

In this work, a novel methodology for the visualization of graphs is presented. Here, graphs are visualized with the purpose of data mining, i.e. the visualization should help the user to understand the concepts and information hidden in the graphs.

The methodology of structure-based visualization comprises three steps: (i) Identification of the graph's structure by means of clustering algorithms, (ii) cluster classification by means of classification functions or case-based classification, and (iii) layout of the clusters and nodes by means of graph drawing methods.

For tabular data sets, a fourth step exists: Node similarities are defined using knowledge acquisition techniques and machine learning.

These four steps are implemented as follows:

1. **Structure Identification:** In this work, graph structures are identified by clustering algorithms. First, existing clustering approaches are described and classified according to two criterions: (i) Optimization criterion and (ii) necessary additional knowledge.
   Then, three new clustering methods are presented:
   - *MajorClust:* MajorClust allows for a fast and reliable clustering of graphs. Theoretical deliberations and practical tests using 6 applications show the high quality of MajorClust's clustering results.
   - *Λ-Optimization*: Λ is a graph-theoretical quality measure for clusterings. Using Λ, features of good clusterings are presented.
   - *Adaptive Clustering*: This method integrates domain-specific knowledge into the clustering process. For this, given exemplary clusterings are analyzed and clustering knowledge is learned.

2. **Structure Classification:** Identified clusters are illustrated by a label, e.g. a network traffic cluster may be identified as a special university institute. For this, two classification approaches are presented: (i) Direct classification and (ii) case-based classification. Machine learning methods are used to abstract the necessary classification knowledge from exemplary classifications.

3. **Structure Envisioning:** In this step, the graphs are laid out, i.e. node positions are computed. First, existing methods are presented and classified according to two criterions: (i) Used quality criterion and (ii) corresponding graph class.
   A novel method, which combines principle component analysis and force-directed drawing, is also presented. This method is applied successfully to several domains
   All graph drawing methods used for this step are extended in order to integrate clustering knowledge.

4. **Creating Graphs from Tabular Data:** In this work, tabular data is treated as a special type of graph. In order to apply the methodology of structure-based visualization to tabular data, similarities between nodes have to be defined. This is done by means of a so-called similarity measure which assesses the similarity between two nodes.
   Since the manual definition of such measures overtaxes the capabilities of most experts, a new methodology for learning of similarity measures, which allows for the application of standard machine learning methods, is described.
   Three knowledge acquisition methods are introduced which implement this methodology: (i) Learning by means of given similarities, (ii) learning by classification, and (iii) learning by examples. All these methods allow for a fast and abstract definition of object similarities.

The visual data mining techniques introduced in this work are applied to 6 domains. For this, the methodology of structure-based visualization and all algorithms have been implemented within the scope of the system StructureMiner. StructureMiner is applied to the following domains:

Network Traffic: Network traffic may be seen as a graph: computers and active network components are modeled by nodes while communication links form the edges. Such graphs are visualized . The following results are worth mentioning as well:
(i) Clusters are classified by a means of a case-based approach.
(ii) The change of traffic over a period of time is shown using a specially developed animation method.

(iii) StructureMiner is used to support the analysis and planing of network topologies.
(iv) The connections between the field of visualization and simulation are explained using a novel network simulation approach.

Configuration Knowledge-bases: Systems for the configuration of technical systems use a so-called knowledge base to store their domain knowledge. This knowledge-base may be seen a graph: technical components form the nodes and causal effects result in directed edges.
Interesting feature are:
(i) Technical subsystem are recognized automatically, i.e. clusters are labeled.
(ii) StructureMiner allows for the choice of different layout methods for different clusters. By analyzing the user's choices, the visualization system learns a function which automatically chooses for future clusters the appropriate layout method.
(iii) Different edge shapes and colors are used to represent different types of edges.

Protein Interaction Graphs: Such graphs model the interactions between proteins and are important in the context of the Human Genome Project. StructureMiner makes instructive insights into these graphs and their structure possible.

Furthermore, MajorClust is used to examine an existing protein taxonomy; it can be proved that this taxonomy groups together such proteins that have only a few interactions with each other.

Fluidic Circuits: This domain is used mainly to present a method which learns a domain-dependent clustering method by analyzing given sample clusterings. Results are given which prove the success of this approach.

Document Management: The visual presentation of large numbers of documents can help to manage and access otherwise incomprehensible document sets. In this work, each document represents a fluidic circuit.

In order to visualize document sets, three steps are necessary:
a) Generation of Document Features: For each document, i.e. fluidic circuit, abstract features are generated automatically. These features give a functional description of fluidic circuits. The reader may note that this domain is therefore an example of the visualization of tabular data sets.
b) Definition of Document Similarities: The knowledge acquisition and machine learning methods described above are successfully applied to the definition of a similarity measure which assess the similarity between document features.
c) Visualization: Using the learned similarity measure from the previous step, a document graph is constructed: documents form nodes; edges are weighted by document similarities. This graph is visualized using the methodology of structure-based visualization.

Visualization of Tabular Data Features: Tabular data is often presented in the form of a table: objects define the rows while object features become the columns. E.g. if objects represent people, reasonable features are name, age, and gender. In this application, StructureMiner is used to visualize dependencies between features. E.g. as an example, a data set consisting of 100,000 people who are described by 481 features, is used.

For this, two steps are necessary:
a) First, the dependencies between features are computed. Depending on the type of feature (cardinal or nominal), different machine learning techniques are used, e.g. regression or $\chi 2$-test. Next a dependency-graph is constructed: Nodes model features; the graph is totally connected. Edges are weighted by the degree of dependency.
b) This graph is visualized using the methodology of structure-based visualization.

The visualization helps the user to understand the dependencies between features. Most of the identified clusters correspond to reasonable clusters, e.g. people living in rural areas or rich neighbourhoods.

In conclusion, it can be said that the new methodology of structure-based visualization is validated in this work by (i) theoretical deliberations and (ii) by applying it to 6 domains. For each step of the methodology, either existing approaches have been used, existing algorithms have been extended, or new methods have been developed.