Graph-Based Tools for ECM Search Result Analysis to Support the Ideation Step

Houcine Dammak[1], Abdellatif Dkhil[1], and Mickaël Gardoni[1,2](✉)

[1] École de Technologie Supérieure, Montréal, Canada
{Houcine.dammak.1,Abdellatif.dkhil.1}@ens.etsmtl.ca,
Mickael.gardoni@etsmtl.ca
[2] Institut National des Sciences Appliquées, Strasbourg, France

Abstract. Enterprise Content Management tool (ECM) is defined as the technologies, tools, and methods used to capture, manage, store, preserve, and deliver content. Using ECM in the ideation step of the innovation process may enhance the creativity of users to create new knowledge. In this paper, we discuss how to access this large amount of information efficiently without overwhelming users which may decrease their creativity. The purpose of this work is to avoid this situation by replacing the classical representation of ECM results with a graphical representation using graph theory. The advantages of this approach for visualizing and analyzing the connection between contents are discussed.

Keywords: Creativity · Knowledge management · Visualization · Content management · Graph theory

1 Introduction

Organizations today have access to an enormous amount of information. This information is present everywhere in the workplace in different formats: documents, communications, databases, etc. and it represents a great asset for any organization. The reuse of information and data-driven management is considered a route to greater efficiency and decision making resulting in improved productivity, profitability and competitiveness [1].

Availability of information is one of the critical success factors for organizations to survive. While trying to overcome this critical success factor, organizations are facing different challenges like the huge volume of data and information that exist in different formats and that is not easy to use.

Since this amount of information is usually non-structured and is present in different locations in the workplace, some information management tools have appeared. From these information management tools, we consider content management tools (ECM: Enterprise Content Management) which have the main goal to manage all the organizations' content.

«ECM is defined as the technologies, tools, and methods used to capture, manage, store, preserve, and deliver content across an enterprise» [2]. It is also defined as a

C. Fortin et al. (Eds.): PLM 2019, IFIP AICT 565, pp. 170–180, 2019.
https://doi.org/10.1007/978-3-030-42250-9_16

collection of strategic resources and capabilities that provides an automated enabling framework for efficient lifecycle management of valuable organization asset, i.e. contents and processes, to carry out required business operations in a collaborative fashion, supports governance and compliance, provides integration within and outside the business boundaries to achieve business intelligence, knowledge management and decision support capabilities with focus on fulfilment of business goals and objectives for competitive advantage [3].

The main ECM steps are [4]:

- Capture: It contains all the activities related to collecting content. It is usually about identifying the content that it wants to capture and all its dimensions. This content could be captured from internal to external databases.
- Organize: It involves indexing, classifying and linking databases together. This step utilizes different techniques like OCR (Optical Character Recognition) and smart templates for indexing (to identify the metadata), workflows for classification based on business rules and ODBC connections to link content with other databases.
- Process: Analyse the content already classified in order to inform decision-makers and other existing management systems.
- Maintain: It is mainly related to the maintenance of the content. How to keep it accessible? How to link it with new content? And for how much time we should keep it?

ECM is used mainly for daily and operational tasks. Usually, the main reasons to implement ECM solutions are: reducing searching times, unifying the presentation or adhering to reporting obligations [5]. ECM has proven its efficiency not only in the industrial world but also in the research world.

In early stages of the innovation process, participants are invited to use their creativity to come up with ideas. Those ideas are usually inspired by the participants' background and own knowledge. Some researchers proposed different tools to support this important stage in the innovation process which is the ideation stage. The combination of teams existing technical knowledge and limited domain-specific knowledge provokes more original and diverse ideas, which confirms there is a creative value in the combination of KDD (knowledge discovery from databases) with teams' existing knowledge [6].

The ideation step in an innovation process is a critical step. Providing tools to participants in ideation sessions may enhance creativity. As mentioned above, ECM is a tool that will be provided to participants of ideation sessions to see its impact on the creativity of an organization. ECM users will have access to valuable information that is structured and easily accessible. This access to information may enhance the creativity of users to create new knowledge. The gap is present, and the benefits expected are really promising but unfortunately until today, in the ECM research, Creativity, Innovation and Knowledge Management have played a minor role.

With the Content Management tool, we are planning to use indexes (Metadata) related to each content to do searches. By using this metadata layer, the participant receives a list of contents as a search result. Most commercial ECM tools present the search results as a list of contents (documents) grouped by the category of the content (Example: Thesis, Article, Product catalog, Invoices…) and details with all the metadata related

(Date, field, author, location, university…). The metadata depends on the category of documents and it is configured at the implementation stage of an ECM solution. So, the participants of an ideation session, depending on the criteria that they select, they receive a list of content as a search result. This list of contents may inspire them in their ideation session to come up with innovative ideas.

2 ECM Contribution in an Ideation Session

2.1 Using ECM in an Ideation Session

As mentioned above, ECM converts non-structured content to structured content using its main functionalities. Then, by providing the ECM as a support tool to participants in an ideation session they have access to all the content indexed and classified. While doing a search in the ECM, participants have the search result in a list format (see Fig. 1). This method offers structured search results for contents to access to the most relevant ones that may support their creativity.

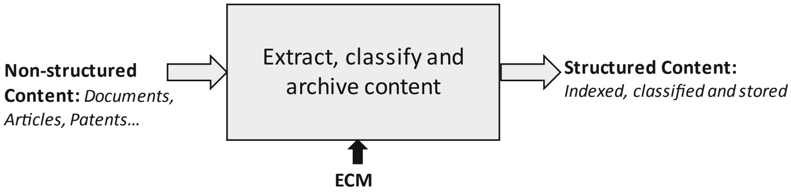

Non-structured Content: Documents, Articles, Patents…
Extract, classify and archive content
Structured Content: Indexed, classified and stored
ECM

Fig. 1. Modeling the ECM system

2.2 Limitation of the Actual Method

The actual representation of search result in the ECM allows users to have access to a list of indexed contents. Usually, organizations have thousands of contents stored in an ECM which may result in a communicative limitation. By having a long list of contents, participants may lose time to demystify this list to understand the indexes and set a priority list. So, the actual method has a limitation in describing the entire structure of contents with multiple relationships. To cope with this challenge, what information should we analyze? and how to display it?

2.3 Proposed Method to Improve Using ECM in an Ideation Session

Some researchers proved that access to large databases of information can overwhelm users, in their innovation process, and tend them to return to known solutions which decrease the creativity [6]. To avoid this situation, instead of presenting the search results to participants as a list of contents with the metadata related, we are proposing a relational analysis between these contents to display them in a graph which may help them.

As mentioned previously, ECM is the platform that stores content and makes it available to users when they need it. So, we are proposing a framework that couples the ECM output and the innovation process at the ideation stage. And this by proposing a relational analysis of the list of contents displayed by the ECM and representing them in a graph (see Fig. 2). The difficulty here is about the important number of analysis properties to identify the right graphs and all the combinations between these analysis properties.

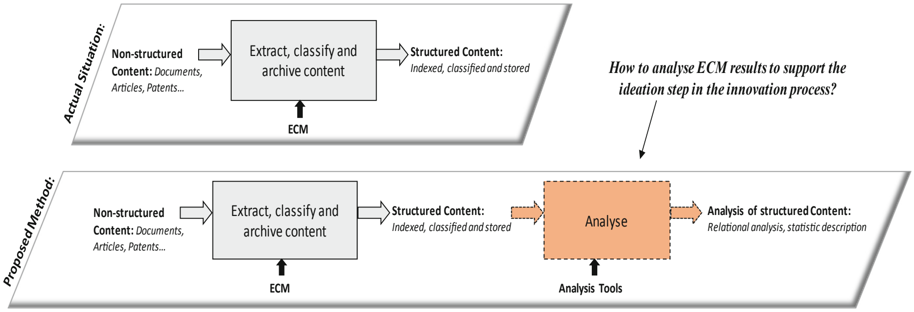

Actual Situation:
Non-structured Content: Documents, Articles, Patents...
Extract, classify and archive content
ECM
Structured Content: Indexed, classified and stored
How to analyse ECM results to support the ideation step in the innovation process?
Proposed Method:
Non-structured Content: Documents, Articles, Patents...
Extract, classify and archive content
ECM
Structured Content: Indexed, classified and stored
Analyse
Analysis Tools
Analysis of structured Content: Relational analysis, statistic description

Fig. 2. Proposed method

2.3.1 Use of Graph Tools to Analyze ECM Search Results

In this study, we propose to use graphs for visualizing and analyzing the result of the ECM application. We define a content graph as a graph in with the nodes represent contents (or indexes), the lines represent the relations, and the labels on the lines represent the nature of the relations. Content Graph is a network with nodes that are connected unidirectionally by links of various relations and are intended to organize the entire relation structure between contents (or indexes). ECM result graph has several noticeable benefits compared to actual search result which is in a list format. The advantages of using graphs will be discussed in the following section.

2.3.2 Advantages of Using Graph Representation

Our methodology consists of using graph theory tools to analyze the ECM results. Using a graph representation have many different advantages. A graph is a standard tool for data visualization. This representation tool permits to minimize the reading and interpretation time. In fact, we propose to use graph representation to better understand the links between contents and to have additional analysis tools. Using a content graph presents several advantages. The content graph simplifies the representation and the relation between contents (or indexes) by using standard representation allowing a comprehensive structure of the organization's content. Graphs have long provided visual languages and have been widely used in many different disciplines as formal representation system [7, 8].

The content graph is a particularly good way of organizing the organization's content residing in the ECM especially when using the search function. Displaying a search result in a graph instead of a list brings a view of the entire content of the search result. Content graph display describes the diversity and the depth of the entire ECM search result without omitting any content residing in the bottom of a list and highlighting the relationship between them. It is essential to participants in an ideation session to understand the relationship between all their organization's content to clarify their thinking and to optimize their creativity exercise.

The other great advantage of this presentation way is the possibility to use many structural analysis tools proposed by the theory of graphs. These tools allow a quantitative analysis of connectivity and relationships between contents [8].

In the following section, we will present some example of analysis tools issued from the graph theory and their utility in the analysis of content graphs.

3 Graph Representation of ECM Search Result Analysis

3.1 Input Data

The ECM tool permits to extract, classify and archive content from Non-structured Content (documents, Articles, Patents…) (see Fig. 1). The output of this system is a list of Structured Content: Indexed, classified and stored. An example of a list of Structured Content is given in the following table.

This example considers that we have two categories of contents (the first one is Thesis) and the first category has three indexes (University, Field and Year) (Table 1). After organizing and structuring the content, ECM delivers it to users in the format of a list. We are using graph tools to change this output to a graph format that highlights the relationships between contents. The goal here is to study the links between all the contents presented in the search result and to present them in a graph format. A common way to represent a graph is the adjacency matrix which is a matrix A(n, n) where n is the order of the graph. An entry (u, v) of the matrix is either 0 if there is no edge between u and v or the weight of the edge (u, v) if it exists (this representation implies that no edge has a weight equal to 0). From the list of Structured Content, it is simple to extract two adjacency matrices which can be transformed into a graph. The first is the content/content matrix A(n, n): n corresponds to the number of content. An entry (i, j) of this matrix A(n, n) indicates the value of the link between content i and content j. The second matrix is index/index matrix. An entry (i, j) of this matrix indicates the value of the link between index i and index j.

Her, and to simplify we use only the adjacency matrix corresponding to content - content link matrix (An entry (i, j) of this matrix indicates the link between content i and content j). From this matrix, a graph called content - graph can be presented.

3.2 Introduction to Graph Theory: Basic Terminology and Notations

A graph G = (V, E) is a mathematical structure often used to define relationships between objects. It consists of a set of vertices V and pairs of vertices connecting them (edges, E). A graph can be directed or undirected. In a directed graph, given the edge e = (u, v), we say that u is the origin of e and v is the destination of e. In undirected graphs, u and v are the endpoints of the edge. An undirected graph (or graph) G = (V, E) consists of a finite set V of vertices, and a set E of unordered pairs of distinct vertices called the edges. We say that vertex v is adjacent to vertex u if there is an edge (u, v). In this paper, the used graph is an undirected graph: this graph corresponds to the content/content matrix A(n, n) [9, 10].

Table 1. Example of a list of structured content

Category				Category 1			Category 2		
				Index	Index	Index	Index	Index	Index
				University (a)	Field (b)	Year (c)	d	e	f
1	Content	c1	Thesis 1	ETS (x)	Innovation (s)	2012 (u)			
1	Content	c2	Thesis 2	McGill (y)	Innovation (s)	2012 (u)			
1	Content	c3	Thesis 3	Concordia (z)	Innovation (s)	2012 (u)			
1	Content	c4	Thesis 4	ETS (x)	Innovation (s)	2012 (u)			
1	Content	c5	Thesis 5	ETS (x)	Innovation (s)	2015 (j)			
1	Content	c6	Thesis 6	ETS (x)	Electrical (r)	2015 (j)			
1	Content	c7	Thesis 7	McGill (y)	Electrical (r)	2015 (u)			
1	Content	c8	Thesis 8	Concordia (z)	Electrical (r)	2015 (u)			
2	Content	c9					q	t	p
2	Content	c10					q	t	p
2	Content	c11					q	t	p
2	Content	c12					q	t	p
2	Content	c13					w	i	l
2	Content	c14					w	o	l
2	Content	c15					w	o	l
2	Content	c16					q	o	k

The order of graph corresponds to the number of nodes and a path in a graph is a sequence of vertices (v0, v1; . . . , vk) such that (vi – 1, vi) is an edge for i = 1, 2, . . . , k. The length of the path is the number of edges, k. A path is simple if all vertices and all the edges are distinct. A path in a graph G is a sequence of vertices such that from each of the vertices there is an edge to the successor vertex. A path is called simple if none of the vertices in the path are repeated. A cycle is a path starting and ending at the same node. A cycle is a path containing at least one edge and for which v0 = vk. A cycle is simple if its vertices (except v0 and vk) are distinct, and all its edges are distinct.

3.3 Limit of the Representation

The graph-based representation permits us to use classical graph theory tools and concepts, but some drawbacks exist. The essential limit is the loss of the details of the link. An edge represents an aggregate index between two contents. Figure 3 shows an example of the aggregation of the link between two content i and j: only one link is established. The weight of this link corresponds to the number of aggregated links.

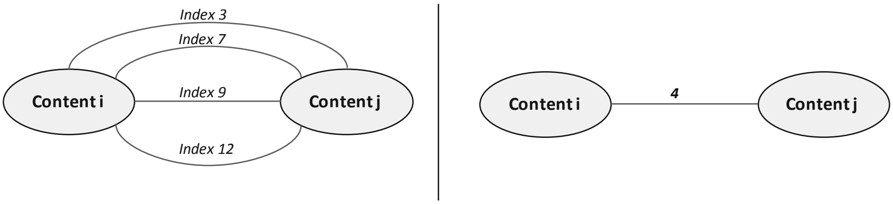


Fig. 3. Example of aggregation

4 Examples of Graph Tools and Utility for ECM Result Analysis

This section aimed to show how classical concepts of graph theory can be applied to analyze ECM search result or to give quantitative information about these results. In our specific context, we present the utility in term of analysis between contents for each concept.

4.1 Biconnected Graphs: Connected Components

Let G = (V, E) be a connected undirected graph, a graph G is connected if, for every pair of nodes v1 and v2, there is a path between nodes v1 and v2. A graph is said to be connected if it can be traveled from any one node to any others by moving along paths of edges. The graph is 2- connected if deletion of any node still keeps it connected; it is 3-connected if it still remains connected with the removal of any two nodes, and so on. It is required that a k-connected have at least k + 1 nodes [9–11].

Notice that unlike strongly connected components of an oriented graph (which form a partition of the vertex set), the biconnected components of a graph form a partition of the edge set.

A graph G = (V, E) is k-connected (k-edge-connected) if at least k vertices (edges) must be deleted to disconnect G. A graph that is 2-connected (3-connected) is also called biconnected (tri-connected).

- Utility: The identification of connected component gives an indication about the clustered nature of ECM search result: each connected component represents an independent cluster. If a graph contains only one connected component, the graph is said to be connected and corresponds to a "one block" draw while a non-connected graph can be drawn in several blocks.

4.2 Articulation Points (or Cut Vertices) and Cut Edge (or Bridge)

Let G = (V, E) be a connected undirected graph. A node in the graph G is called cut point (or articulation point) if its removal disconnects a graph, i.e. increases the number of components. Also, it makes some points unreachable from some other [9]. Articulation Point (or Cut Vertex) correspond to any vertex whose removal (together with the removal of any incident edges) results in a disconnected graph.

In graph G, a Bridge or cut edge is an edge whose removal results in a disconnected graph. An edge is a bridge if its removal results in disconnected sub-graph. A bridge is an edge such that the graph containing the edge has fewer components than the sub-graph that is obtained after the edge is removed.

Figure 4 shows an example of cut vertex and cut edge. cut the bridge disconnects the graph and forms disconnect subgraph (cluster).

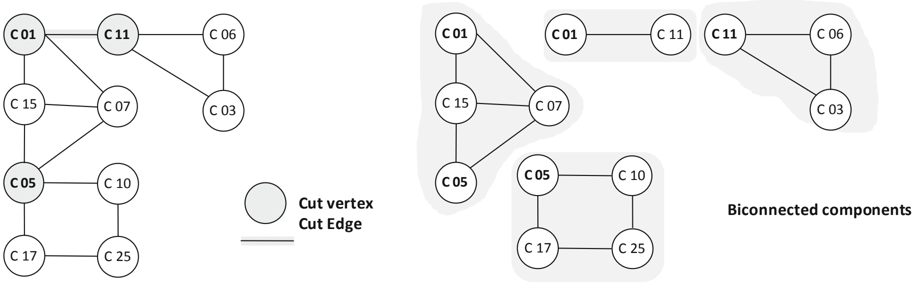


Fig. 4. Example of cut edge and cut vertex

A graph is biconnected if it contains no articulation points. (In general, a graph is k-connected, if k vertices must be removed to disconnect the graph). The concept of a cut point can be extended from a single node to a set of nodes necessary to keep graph connected. Those nodes are referred to as a cut-set. A node cut-set is a subset of the nodes of a graph, whose removal (simultaneously removing all edges adjacent to those nodes) makes the graph no longer connected. If the set is of size k, then it is called a k-node cut, denoted by k(G). That is, the K(G) of a graph is the minimum number of the nodes that must be removed to make the graph G disconnected.

- Utility: Biconnected graphs, articulation points, bridge are of great interest in the analysis of d content graph because these are the "critical" points, whose failure will result in the network becoming disconnected. Also, the biconnected components of a graph are the equivalence classes (see Sect. 4.1).

4.3 Clustering

Clustering is a process of finding such groups based on chosen semantics. According to this semantics, the current clustering approaches can be roughly classified into two categories: content-based clustering and structured based clustering. Content-based uses semantic aspects of data such a category labels, while structure-based clustering takes advantage of structural information about data. Moreover, structured-based clustering is domain-independent so that it is suitable for graph visualization.

In order to cluster a graph, a metric of a node in the graph is required to quantify its features. Based in this metric, existing approaches of partitioning graphs [10, 12] can be loosely divided into the following groups: connectivity based partitions, which use standard concepts from graph theory, distance partitions from selected subsets, Neighbourhood based partitions, and other approaches.

There are several ways how to rearrange a given matrix (correspond to the graph) determine an ordering or permutation of its rows and columns. To get some insight into its structure:

- Utility: The goal of clustering is to reduce a large, potentially incoherent network to a smaller comprehensible structure that can be interpreted more readily. A clustered graph can greatly reduce visual complexity by replacing a set of nodes in a cluster with an abstract node. Clustering, as an empirical procedure, is based on the idea that units in a network can be grouped according to the extent to which they are equivalent, according to some meaningful definition of equivalence.

4.4 Graph Node Groups (Collapse/Contraction)

Collapsing graph is an alternative way to reduce visual complexity. Collapsing means removing from the visualization the nodes that are connected to one node or to a group of nodes. Any number of nodes can be collapsed into a single synthetic node: collapse set of nodes and expand it when needed.

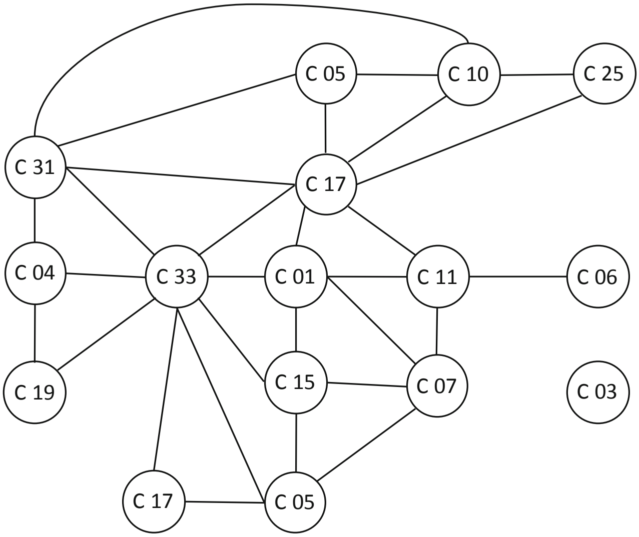
Such a synthetic node contains a user-provided text instead of normal disassembly listing.

- Utility: If a graph is too large to fit on the screen, groups of related nodes are (clustered) collapsed into super-nodes. The users see a “summary” of the graph, namely the super-nodes and super-edges between the super-nodes. Some clusters may be shown in more detail than others. The process collapsing involves discovering groups in the data. In the case of graph visualizing collapsing nodes Groups can be un-collapsed to display the original node content.

4.5 The Degree of a Vertex and Adjacency List

A graph consists of vertices and edges connecting these vertices. The degree of a vertex i is the number of edges incident with it, except that a loop at a vertex contributes twice to the degree of that vertex. The degree of the vertex v is denoted by deg(v) and calculate from the adjacency or the neighborhood of vertex. Two vertices u and v are called adjacent or neighbors if u and v are endpoints of an edge e of G = (V, E). The degree of a vertex represents the number of edges incident to that vertex. Figure 5 shows an example of a graph and the degree of each vertex.

- Utility: To analyze a graph it is important to look at the degree of a vertex. The degree of vertex informs the Criticality of contents. Critical contents are those that have a significant number of links to other contents. In this example, content C33 correspond to 14% of the total number of links.



V	deg(v)
C 31	5
C 05	3
C 10	3
C 25	2
C 17	7
C 04	3
C 33	8
C 01	5
C 11	4
C 06	1
C 03	0
C 07	4
C 15	4
C 19	2
C 05	4
C 17	2

Fig. 5. The degree of a vertex

4.6 Singleton

The singleton graph is the graph consisting of a single isolated node with no edges. It is, therefore, the empty graph on one node. In Fig. 4, the vertex D03 is a singleton.

- Utility: As known, there are two types of innovation: Incremental and Disruptive (or radical). Incremental innovation is a series of small improvements or upgrades made to a company's existing products, services, processes or methods. In the other hand, disruptive innovation is an invention that changes radically an existing product, services, process or method. We believe that a singleton in our content graph may bring a disruptive idea to participants in the ideation session.

5 Conclusion

This document intersects with the improvement of the use of the ECM result in the ideation phase. We propose to use graph theory to facilitate access to the analysis results.

The advantage of using the visualization of the results in the form of a graph was presented. The tools presented in this paper permit to analyze ECM results and to give indications about the relation between the used contents.

These tools aimed to highlight some particularities which are useful in the ideation session. Future research should consider some of these suggestions to further extend this line of research. Also, it would be interesting to explore other analysis properties.

References

1. Hicks, B.J., Culley, S.J., Allen, R.D., Mullineux, G.: A framework for the requirements of capturing, storing and reusing information and knowledge in engineering design. Int. J. Inf. Manag. **22**, 263–280 (2002)
2. AIIM: Association for Information and Image Management. www.aiim.org
3. Usman, M., Muzaffar, A.W., Rauf, A.: Enterprise content management (ECM): needs, challenges and recommendations. In: 2009 2nd IEEE International Conference on Computer Science and Information Technology, 8–11 August 2009, pp. 283–289 (2009)
4. Nonaka, I.: The Knowledge-Creating Company (Harvard Business Review Classics). Harvard Business School Press, Boston (2008)
5. Brocke, J.V., Seidel, S., Simons, A.: Bridging the gap between enterprise content management and creativity: a research framework. In: 2010 43rd Hawaii International Conference on System Sciences, 5–8 January 2010, pp. 1–10 (2010)
6. Escandon-Quintanilla, M.L.: Effects of data exploration and use of data mining tools to extract knowledge from databases (KDD) in early stages of the engineering design process (EDP). Doctoral dissertation, École de Technologie supérieure (2017)
7. Bang-Jensen, J., Gutin, G.Z.: Digraphs: Theory, Algorithms and Applications. Springer, Heidelberg (2008)
8. Herman, I., Melançon, G., Marshall, M.S.: Graph visualization and navigation in information visualization: a survey. IEEE Trans. Vis. Comput. Graph. **6**(1), 24–43 (2000)
9. Eades, P., Tamassia, R.: Algorithms for drawing graphs: an annotated bibliography. Brown University (1988)
10. Harel, D., Koren, Y.: Graph drawing by high-dimensional embedding. In: Goodrich, M.T., Kobourov, S.G. (eds.) GD 2002. LNCS, vol. 2528, pp. 207–219. Springer, Heidelberg (2002). https://doi.org/10.1007/3-540-36151-0_20
11. Hossain, M.I., Rahman, M.S.: Good spanning trees in graph drawing. Theor. Comput. Sci. **607**, 149–165 (2015)
12. Everitt, B., Landau, S., Leese, M., Stahl, D.: Cluster Analysis. Wiley, Chichester (2011)