Infomat Manual

for version 080829

Magnus Rosell

August 29, 2008

Contents

Chapter 1

Introduction

Infomat is both a processing tool and a visualization tool. This manual will (when it is complete) deal both. The visualization will be covered in Chapter 2 and the (non visual) processing will be covered in Chapter 3.

Please note! that this is work in progress. It is definetly not complete and may in part be out of date as I devlop the program all the time. Still, I hope it will come to some use.

Further information can be found in the `readme.txt` file, the javadoc of the program, which can be found in the `doc` subdirectory, and on the Infomat website:

`http://www.csc.kth.se/tcs/projects/infomat/infomat/`

1.1 Infomat Basics

Infomat deals with objects that are called *IObjects*. Each IObject has a string and an id number that uniquely identifies it. It has also, when applicable, a reference to a location where the actual object is stored (like a actual text file). In this manual they often will be called *objects*, for short.

Several IObjects can be stored in an *IObjectGroup*, and several IObject-Groups consitutes an *IObjectGrouping*. Through this manual these are also called groups and groupings for short. Right now each IObject can belong to only one IObjectGroup in every IObjectGrouping.

The main data structure in Infomat is a matrix, called an *IMatrix*. It is an implementation of a sparse matrix[1]. The objects along the axes of the matrix, rows and columns, corresonds to IObjects. Each axes has a special IObjectGroup called an *IObjectSet*. The IObjectSet also keeps track of all IObjectGroupings of it. An IObjectGrouping can only contain IObjects from one IObjectSet.

The IMatrix stores several *IMatrixCells* which holds information of the relation between two IObjects, one from each IObjectSet. The basic information is a count, and a derived information is called a weight.

[1] I am slowly developing a dense matrix structure, which might be useful sometimes. However, for now all matrixes are handled as sparse. As the intended use of Infomat is Information Retrieval it is not a problem. For the GUI to be really useful the objects along the axes of the matrix has to be interpretable. When they are, the matrix is usually sparse.

For a typical Information Retrieval scenario the row IObjects may consitute texts, with titles and locations in the file system, and the columns words that appear in the texts. For each word that appear in a particular text an IMatrixCell with the number of appearances is stored as the count. The weight of the IMatrixCell can be calculated through a weighting scheme. An IObjectGrouping of the texts (rows) could be a clustering or a categorization of the texts.

Any information stored in a matrix may be investigated using Infomat.

Chapter 2

Infomat as a Visualization and Exploration Tool

In this chapter Infomat as a visualization tool will be described. It allows you to display a matrix, and group, order and alter it. You may do this along the rows or columns. This chapter describes the GUI in an order that follow the layout. It should be considered as a reference.

The last two sections are a bit different. Section 2.9 describes the most important concept of the GUI and Section 2.10 describes the example matrix that is bundled with the program.

There is no undo function. Save your work!

2.1 Interface overview

Figure 2.1 shows the interface. The **main window** is divided into four sections. At the top is the **menu**, below that the **toolbar** and under that the **grouping panel**. The **main view** is the fourth section. There are several other windows that appear in certain situations. From the beginning the **Overview** is shown.

In the following sections the main window sections and several other windows will be described briefely. Section 2.8 describes a few GUI components that appear in several places. Finally, Section 2.10 describes a small example.

2.2 Main View and the Overview

Infomat stores a matrix. The main view and the overview display the matrix. The opacity of the pixels are proportional to the weight of the matrix elements they represent.

When you load a matrix, see Section 2.3.1, the whole matrix is displayed, but many operations result in a partial view. Which part is shown is decided thorugh the *Grouping Panel*, see Section 2.6.1. The Main view, further, may be zoomed in on any part of the partial view. The Overview always displays all of it, and indicate by a rectangle what part the main view shows.

The main view and the overview display a part of the matrix.

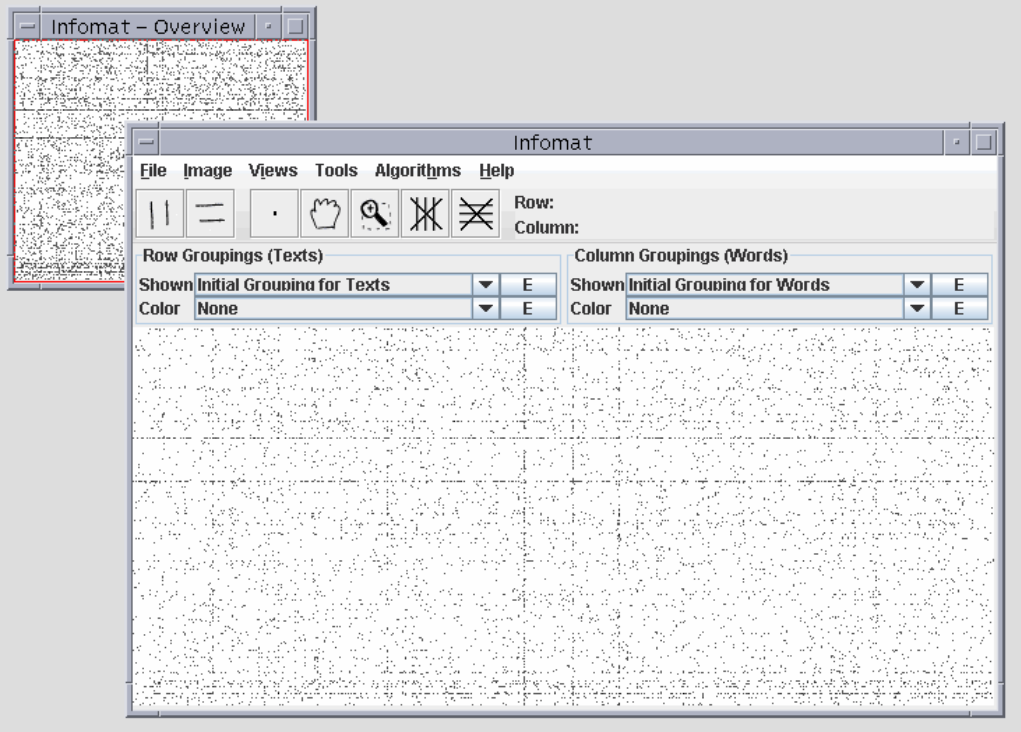

Figure 2.1: Infomat.

2.3 Menu and Toolbar

This sections contains a short account for the available menu options. As the toolbar contains convenient short cuts to some of the options it is described here as well.

When a matrix is loaded and the mouse pointer is moving over the main view the row and column objects it is currently pointing to is displayed in the toolbar. (However, it is only the objects in the topmost row and the leftmost column of the current pixel. To see more use the Pixel View window.)

The following subsections describes the content of the menus.

2.3.1 File

In the file meny you can save and load matrix files. It is also possible to load a “token file”, which is a single file conataining several texts. Look at the exampel, Section 2.10, for the format.

Under *Other* you can load two other formats. The "corpus files" and "document grouping files" is a connection to an older clustering program I wrote. You are probably better off not looking in to them... :)

It is also possible to save the picture in the main view as .png-file.

The Infomat Properties are som fundamental settings for the program. They are displayed and altered through the Properties-GUI which recurs for several settings through the program.

2.3.1.1 File Formats

The xml-formats are quite straight forward. You should be able to figure them out by looking at the examples, see the "readme.txt".

2.3.2 Image-menu and Toolbar

The toolbar is divied into two sections, with two and five buttons. The five first menu options on the Image menu corresponds to the five icons in the second button section:

Pixel selection When the mouse is clicked on a pixel in the main view information on it is displayed in the Pixel View window, see Section 2.3.3.

Drag For moving the selected zoom area.

Zoom selection By clicking, dragging and releasing the mouse within the view that area is zoomed in.

Delete rows Click, drag and release to remove rows.

Delete columns Click, drag and release to remove columns.

The following two menu options corresponds to the leftmost two icons. They toggle the group separators on/off. The last two options in the Image menu toggles the guide lines (that helps with positions) on and off and resets the zoom entirely.

All Image functions work in both the main view and the overview.

2.3.3 Views

The different Views are the main ways to get information. The options in the Views menu are all toggle options: activating/deactivating the corresponding view.

Pixel view The pixel view shows information on the pixel the mouse pointer is currently pointing on and very much more. It is described in Section 2.4.

Similarity View In the similarity window you can change what similarity measure is used for both rows and columns. You may also change the Properties of the chosen similarity. The choosen similarities are applied whenever appropriate: for many clustering algorithms and when sorting IObjects accoriding to similarity in different ways.

Overview The overview window shows the entire matrix and indicates which part is currently visible in the main view.

Grouping The grouping panel with all its functions is described in its own section, Section 2.6.1.

Tooolbar The toolbar is described in Section 2.3.2 on the Image menu.

2.3.4 Tools

There are several tools:

Evaluation Choose grouping to evaluate (and if you want to make an external evaluation a reference grouping) and press "Evaluate". The measures can be saved and loaded in an xml-format.

Matrix Summary gives some basic matrix information.

Export to Text exports the currently selected grouping(s).

Stoplist is a rather complex tool that is descirbed in Section 2.5.

Purge matrix removes all objects that are not displayed in the overview. If you for instance have deleted certain uniteresting objects from a grouping this function removes them from the matrix and from all other groupings. Purge matrix is applied the moment you chose it from the menu, whithot any options window.

Invert matrix speaks for itself. As for Purge matrix, this is applied directly.

2.3.5 Algorithms

The options in the algorithm menu are:

Clustering Algorithms Here you can choose between several clustering algorithms. See Section 2.7.

Filter Matrix The Filter Matrix algorithm is straight forward: alter the Properties and hit the apply-button.

Weight Matrix In the Weighting window you can choose between different weighting schemes and alter their properties. When you are satisfied, hit the apply-button. The weightings considered the rows to be the objects, and the columns the representation.

Some things in the properties need explanations:

tf accoring to Croft 1983

$$\mathrm{tf}_{i,j} \quad = \quad c_1 + (1 - c_1)\frac{n_{i,j}}{\max_i n_{i,j}} \tag{2.1}$$

idf accoring to Croft and Harper 1979

$$\mathrm{idf}_i \quad = \quad c_2 + \log \frac{n - n_{\mathrm{word}(i)}}{n_{\mathrm{word}(i)}} \tag{2.2}$$

where $n_{i,j}$ is the number of times word i appears in document j ($\max_i n_{i,j}$ is the number of times the most frequent word in text j appears). In the properties c_1 is called *(Local row/global column) weight importance factor* and c_2 *Global column weight belief factor.*

Cluster Sorter The cluster sorter is work in progress. It allows you to sort the clusters of the current clustering of the rows in order of their value for a evaluation measure. Choose the reference grouping, choose "ascending" or "descending" and hit "Sort by value".

2.3.6 Help

Try them! :)

2.4 Pixel View

Through the Pixel View you get textual information about the matrix as ordered by the current groupings. This is a rather complex tool and is described in some detail here.

When you open it the first time it has just two panels. From the top the *main* and *current* panels. When the mouse pointer points to a particaluar pixel the current panel displays all the matrix elements that the pixel represents.

2.4.1 The View button and the Current Lists

The *View*-button in the main panel lets you choose between elements, rows and columns. For elements the *Elements*-tab in the current panel displays all the matrix elements that are represented by the pixel the mouse pointer points to. They are presented as pairs, like: (row-object, column-object) followed by a value, the weight of the matrix element. If you choose the *Rows*(*Columns*)-tab the row(column)-objects for the matrix elements are presented with the weight of the corresponding elements.

When the *View*-button is set to *Rows*(*Columns*) the Elements-tab does not show anything. The Rows(Columns)-tab shows the row(column)-objects associated with the picture row(column).

The selection (next section) is affected in the same way by the View-button.

2.4.2 Selection

The *Select*-button in the main panel opens (and closes) two panels: the *select* and *selected* panels. These panels allow you to study some objects more thouroghly.

To select anything pixel selection has to be on. See Section 2.3.2.

The *Select*-button in the *select* panel lets you choose between *Single* and *All in area*. If you click the mouse on a pixel when set to *Single*, the objects in the current lists are stored in the selected lists. When set to *All in area* you can select several objects by a click-drag-release procedure.

The objects stored in the selected lists are the recently selected. They stay there when you move the mouse.

2.4.3 Gathering

The *Gather*-button in the main panel opens (and closes) two panels: the *Copy selected* and *Gathered* panels. The gathered panel allow you to store the addition of several sets of selected objects (as described in the previous Section, 2.4.2).

The *All*, *Elements*, *Rows* and *Columns* buttons in the *Copy selected* panel adds the corresponding list of the selected objects.

The *How* button lets you choose between three things. When it is set to *Accumulate* if an object is already in the gathered list its value is increased with the value it has in the selected list. When it is set to *Add* the latest value for the object is stored, and when it is set to *Set* all previous objects are removed and the selected ones are added.

2.4.4 More

The *More*-button in the main panel opens (and closes) three panels: the *Remove Gathered*, *Sort Gathered to Selected*, and *Select Gathered* panels.

The *Remove Gathered* panels allow you to remove the gathered objects from the matrix. The elements are removed from the matrix (it thus affects all groupings), while the rows and columns are only removed from the current shown groupings. To remove these from the matrix and all groupings you need to use the *purge matrix* opition in the *Tools* menu, see Section 2.3.4.

The *Sort Gathered to Selected* panel allows you to sort the gathered row(column)-objects in order of similarity to the all row(column)-objects of the selected lists:

RowRow The similarity of each gathered row to all the selected row. The row similarity measure (see Section 2.3.3) is used to extract the representation and calculate the similarity.

RowCol The similarity of each gathered row to all the selected columns considered as a representation, using the row similarity measure.

ColCol The similarity of each gathered column to all the selected columns, using the column similarity measure.

ColRow The similarity of each gathered column to all the selected rows considered as a representation, using the column similarity measure.

The *Select Gathered* panel lets you move the gathered objects to the selected panel. There are four straight forward buttons: *All*, *Elements*, *Rows* and *Columns* sets the selected objects, overwriting the previus selected objects. The *El for RC* extracts the matrix elements that intersect with the rows and columns of the gathered objects and sets them as selected objects. The *RC for El* does the opposite.

The last row of the *Select Gathered* panel lets you extract the representation for the objects. With the *C for R* button you set the columns of the selected list to the objects that represents the rows of the gathered list according to the row similarity. The *C for R* button uses the column similarity analogously.

2.5 Stoplist

The stoplist tool is an implementation of the common notion of a stoplist in information retrieval. It can do a little bit more though.

The stoplist window has four panels. The leftmost shows several Properties, that might be altered. The middel panel allows you to load and save a list of ordinary strings from/to a simple text file.

The rightmost third of the stoplist window consists of two panels. The top panel is a list of IObject:s that can be removed (stoppped). These might be loaded from an xml file (and saved as well).

The button *From Strings to IO* allows you to convert the strings into IObject:s that can be removed from the matrix. Only IObject:s that exist in the matrix are generated. IObjects may be converted into a list of strings using the *From IO to Strings* button.

In rightmost bottom panel *Main* you choose which matrix dimension that is considered. The *Apply* button removes the IObject:s currently in the IObject list from the matrix and all groupings.

Using the "From IO to Strings" button you can save any list of objects in a simple text format.

2.6 Groupings and Groups

Infomat stores a matrix. It is displyed in the main view and the overview in order of a row and a column grouping. A grouping consists of one or several groups, which together contain all or some of the IObjects in the matrix. This sections describes how the groupings and groups are managed.

2.6.1 Grouping Panel

Through the grouping panel all handling of the groupings is deviced. It is divided into two sections, one for rows and one for columns. They work similarily.

The topmost drop down menu displays the currently selected grouping. When you choose the grouping here the order of the objects along the dimension (rows or columns) changes.

The bottom drop down menu selects the coloring grouping. For the rows this leads to a coloring of the pixels, and for the columns a coloring of the background columns. The pixels are averaged over the matrix elements they represents, while the column coloring is averaged over the entire columns.

When the *E*-button beside each drop down menu is pressed a grouping edit window is displayed. It is described in the next Section, 2.6.2.

2.6.2 Grouping Edit Window

The grouping edit window looks a little different depending on which of the four groupings it concerns. They all have the following sections:

Name panel Here the name of the grouping is displayed. You can alter it.

Groups panel Here all the groups are displayed. For each group you can alter the name and press the *E*-button, which opens up a group edit window. It is described in Section 2.6.3.

Reordering panel By changing the order of the numbers in the text filed and pressing the *Apply* button you can change the order of the groups in the grouping. If you leave a group out it is deleted - a very convenient way to remove one or more groups.

File panel Here you can load and save groupings. For either to work there has to be a matrix loaded.

For coloring groupings you can change the color of each group in the groups panel. The change does not take effect until you press the *Apply* button in the coloring panel which (for coloring groupings) is located between the reordering panel and the file panel. There you can also reset the coloring to the default colors.

The opacity of the pixels can be altered in the "row show" grouping edit window and the opacity of the column coloring in the "column coloring" grouping settings window. By default the column coloring opacity has a lower range than the pixel opacity.

2.6.3 Group Edit Window

The group edit window have the following sections:

Info panel Here the name of the grouping is displayed. You can alter it.

Main panel Here, you can apply any changes you make in the list panel to the actual group, using the *Apply List Order* button.

List When you open a group edit window the list panel contains all the IObjects in the group. You can alter it in many ways, using the list manipulations. The similarity that is used is the row or column similarity from the Similarity View, see Section 2.3.3. For the manipulations to affect the group you have to press the *Apply List Order* button.

2.7 Clustering Algorithms

There is a Clustering Algorithm Window. In it you can decide if you want to cluster rows or columns. You choose algorithm in a combo box. The algorithms all have some properties that can be altered, like for instance the number of clusters. The algorithm window explains these properties rather well.

K-Means K-Means clustering

Bisecting K-Means Bisecting K-Means clustering

Relative Clustering An algorithm that cluster the columns (or rows) relative the rows (columns). The column objects that have the higest weight in the first row cluster is assembled into a first column cluster, and so on.

Random Clustering Just what it sounds like.

Location Grouper constructs a grouping based on the location of the objects in the file system, if this information is available.

The clustering algorithms are applied to the whole matrix, not just the part that is displayed at the moment!

2.8 Standard Components

The GUI makes use of some standard components that appear in several places. This section describes some of their functions in more detail.

2.8.1 Buttons

Most buttons have direct effect. There however, are several alternating buttons that only sets the contexts for actions. The typical example the *Choose rows or columns:* button in the Clustering Algorithms window. It alternates between the words *Rows* and *Columns* when you press it, the visible being what you have choosen. Most alternating buttons has a leading text ending with a colon and it should be rather obvious from the context.

2.8.2 Properties

A lot of functions could be applied in several different versions. Instead of presenting all of them separately they have properties that you can alter. These properties can be saved in an xml format and recalled. There is also a default setting.

The properties gui is easy to understand. Each property has a value that is displayed. You can alter it by typing in a new value and hitting enter. If your new value is allowed the value of the property changes, if it is not it does not.

Some property values are completely open, you can type in anything you want. If you type something inapproprate here (like for instance -5 clusters in a clustering algorithm) and apply the function Infomat throws java Exceptions.

2.8.3 Lists

Lists appear in several functions. They can display IObjects, and IMatrix-Cells and simple java Strings. Lists have many functions. Depending on the

context not all of them are available. It is for instance not possible to load and save IMatrixCells.

The list gui consists of two parts: the list of objects and above that a few functions. The objects are presented in a textual way, often accompanied by a value indicating their order. If the name of the object appears like a button you may open the object in a simple viewer by clicking it. The objects also have a checkbox that you may tick.

The function part of the gui has at most three rows, from the top:

File The file row lets you load and save the list of objects. The loading is usually restricted by a IObjectGroup, meaning that you will only actually load those that is in that particular group. For the Group Edit window it is the corresponing IObjectGroup. In most other cases it is the IObjectSet of the IMatrix corresponding to the rows or columns depending on the context.

Sel The selection row allows you to handle selected objects. The *Sel* button selects all objects, the *Desel* button deselects all objects, the *Rm* button removes all currently selected objects, and the *Inf* button inverses the selection.

Order Here you can reorder the objects in the list. There are several possible orderings. You choose between them in the combo box and applied them on the objects by pressing the *Apply* button.

This has no other effect than the ordering in the list. For an ordering to have effect on anything else you have to do more. On a group for instance you have to hit the *Apply List Order* button, see Section 2.6.3.

The two first uses the similarity measure along the rows or columns depending on the context. These are only available in some of the lists.

Sim. to Sel. Sorts all objects in order of similarity to the selected objects.

Similarity Sorts the objects in order of similarity to all the objects in the list.

Literal Sorts the object in literal order.

Random Makes a random permutation of the objects

Invert Inverts the order of the objects.

Original When the list is displayed for the first time it has a particular order. Through this you can revert to it. There is one exception. When hit the *Apply List Order* button in the *Group Edit Window* the new order is set as the original.

2.9 The Matrix – Grouping Concept

Infomat is quite a complex tool. The single most important thing to keep in mind when using it is that the view presents the matrix through a row and a column grouping. The matrix may contain several objects that are not

visable. Some functions work on the visual groupings and some work directly on the matrix behind them. This section describes some implications of this fact.

Each grouping is a view of the matrix. Use the purge matrix option in the tools menu to force the matrix to contain only the objects in the current row and column groupings. The objects are also removed from all other groupings similtaneously.

Some functions work on the groupings and some work on the matrix directly. When you remove matrix elements you always remove them directly from the matrix. Row or column objects, on the other hand, are removed either from the current grouping or the matrix directly, depending on the tool you use.

A list of functions that remove row and column objects from the groupings and not from the matrix:

- Through the toolbar and image menu.
- Through the Pixel View window.
- Through the Group View window (after you have pressed the *Apply List Order* button).

When you have removed objects and/or matrix elements (and purged the matrix), remeber to weight the matrix again, using the Weight Matrix function in the algorithm menu.

2.10 Example

In the directory `/Infomat/example/` you find a few files to start with. Read more in the `readme.txt` file.

Chapter 3

Infomat as a Processing Tool

This chapter will explain (some) of the possiblities with Infomat when it is not used with the graphical userinterface as a visualization tool. Its parts are described in the javadoc in the directory `/Infomat/doc/`.

For the time being, look at the example in the `readme.txt` file: the `ExampleClusterer` class.