Use cases of Impure, an information interface

Santiago Ortiz*
Bestiario

Víctor Pascual Cid†
Bestiario

ABSTRACT

Impure is a visual programming language that enables users to gather, manipulate and visualize many types of data. Using a node link approach where nodes are data operators or visualizations and links represent data flow, users can construct simple and complex mashups to analyze and understand data from many different sources. We present two use cases of Impure based on user feedback, with which we are working to upgrade the system.

Index Terms: H.5.2 [Information Systems]: Information Interfaces and Presentation—User Interfaces; H.4.m [Information Systems]: Information Systems Applications—Miscellaneous

Impure is a visual programming language intended to empower and simplify the exploitation of data.The system enables users to gather information from different sources such as user owned data, diverse internet feeds (including social media data), real time or historical financial information, images, news, search queries results, to name just a few. A wide range of data operators are available in order to mix different data sources that can finally be explored through a set of interactive visualizations.

The power of Impure relies on the generation of data mashups enabling both easy access to data sources and the possibility to explore them using interactive visualizations. Such visualization can, in turn, be part of the analysis process generating outputs that can feed other visualizations or operators to generate richer data. Therefore, Impure constitutes a tool that enables the generation of workspaces that facilitate the exploration and understanding of data.

Today, the Internet is the largest information repository in the world. The vast amount of historical information that it contains, combined with the widespread use of social networks, which also generates data in real time, provide an invaluable resource, an understanding of which can provide many benefits to any kind of user, organization or company. However, these information consumers do not usually have the programming experience to collect and process the available data. In an effort to simplify this process, the main goal of Impure is to assist non-expert users in the manipulation and understanding of data through a set of visual elements that help generate data flows.

1 RELATED WORK

Visual programming has been an object of broad research. Surprisingly, some of the most interesting work has been carried out in the music field, where systems such as Pure Data [4] or the recent OpenMusic [3] represent successful examples of this kind of technology.

Due to the information revolution sparked off by the Internet, new systems are now emerging for combining data from different sources, such as Yahoo Pipes [2] or the Wishful Search system [5]. However, little work has been done in providing a system that enables users to generate data mashups that can also be manipulated with interactive visualizations.

On the other hand, Many Eyes [6] and Tableau Software [1] are examples of systems the provide powerful interactive visualizations that facilitate exploration and understanding of data. However, in such systems, visualizations represent the end point of the analyses. Contrary to that approach, Impure visualizations can be understood as visual operators that also help filter, focus and manipulate data.

2 IMPURE MODULES

The Impure visual interface is based on a classic node link approximation where nodes represent modules that can be interconnected.

Modules may have input parameters (receptors), some of them optional, and output parameters (emitters). The event-based development enables the system to be aware of possible changes in any of the modules, running the whole cycle again to update all of them.

Impure modules are organized in 5 different categories:

- *Data structures*: such modules hold data coming from a data source (i.e. Number, String, List).
- *Operators*: they have one or more receptors, that enable the user to perform a specific operation (i.e. addition or substraction) which returns the result in a single emitter.
- *Controls*: they act as dynamic filters (i.e. interval selectors).
- *Visualizators*: they receive data structures from operators or controls and provide interactive visualizations of the data. They usually return emitters on selected visual objects, that can be used as input into another module.
- *APIs*: modules that allow immediate communication with many sources of information on the Internet. Among the most common are: google search, twitter search, twitter word historical behavior, market data, flickr search, flickr sets load, delicious account data load, ebay items information, dictionary definitions, semantic expansion and so forth.

All modules are listed and tagged according to their functionalities in a menu that enables the user to search for the most relevant actions to perform given a specific type of data. Such modules can be dragged to the workspace, and connected among them.

Figure 1 shows an example of a Impure workspace where different data files are loaded and visualized.

3 USE CASES

We have collected information on our tool's impact at two major international companies (Repsol and Havas Media). Due to the learning curve of the system, at least 3 users from each company have been specifically trained to generate data workflows, while the rest of the users (up to an average of 30) have the task of feeding such workflows with data regarding specific use cases and analyze them. We provided workflow producers with a set of examples that were in situ explained and dissected in order to train them with the tool.

Below, we present two case studies from real users who agreed to share the workspaces they created with Impure as well as a description of the main impact that our tool produced in their working practices.

*e-mail: santiago@bestiario.org
†e-mail:victor@bestiario.org

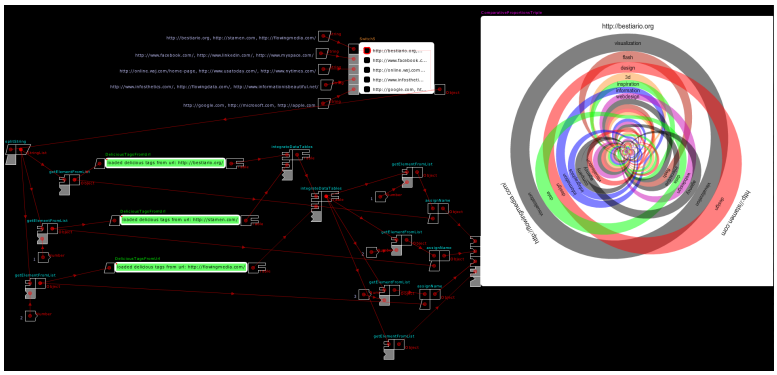

Figure 1: Different modules with data sources and a visualization in Impure allow to compare the tags existing in the Delicious service to describe the URLs from three different companies. Top modules are data structures which allow the selection of different strings used in a search on the Delicious API (green modules on the left). Central modules are controls that manipulate and aggregate the number of results received, which are finally used as input in a visualization module.

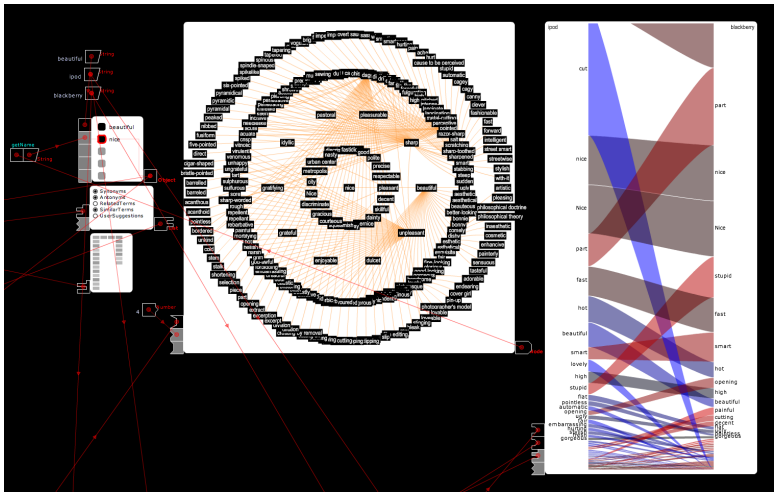

Figure 2: Workflow 1: Example of the workspace that enables the comparison between two different brand names. Left visualization shows words from the Big Huge Thesaurus that feed the module that launches queries.

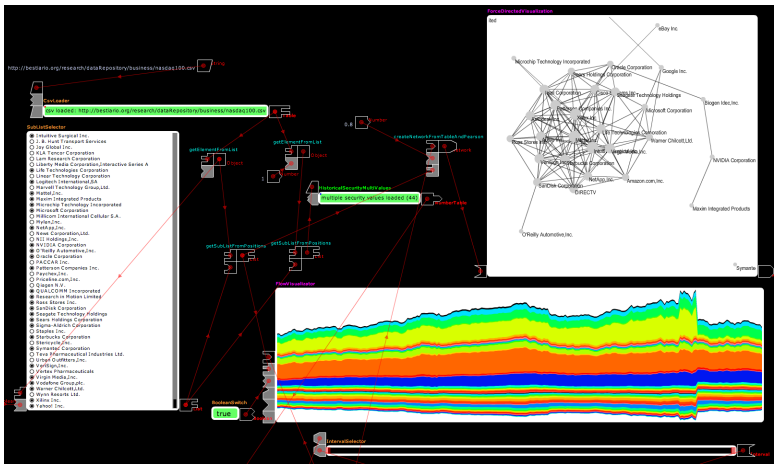

Figure 3: The list on the left enables to select the tickers to be tracked. Any change in the selection in that list turns into an update the two final visualizations: a graph showing tickers relationships and a stacked plot to reveal their main trends. The slider control in the bottom helps the selection of a data range that also affects both visualizations.

3.1 Study of brand perception

The purpose of the use case is to understand how customers and potential customers refer to different products and brands and how this relates to sales behavior.

The Havas Media lab has developed multiple Impure workspaces with interesting results. One of the most relevant, and which gave a larger number of results, is a system that automatically searches for phrases on the Internet which help compare two brands. In this case, Ipod and Blackberry where compared with sentences such as 'iPod is X' or 'Blackberry is X', where X is a word belonging to a list of adjectives. The analysts idea was to quantify the number of occurrences on the Internet of these two searches, in order to compare them. However, one of the first problems that emerged was the need to create a non-arbitrary list, based on terms that are actually used in forums and social networks. In order to solve this problem, the analysts did a search for which adjectives are associated with the brands through a semantic network based on a specific adjective, using information from the api-module which searches for related words (synomyms, antonyms, related) based on the api Big Huge Thesaurus[1]. Thus the analysts created a visual and interactive space using a circular layout with words that affect the final search.

Figure2 shows part of this workspace, including the semantic network for exploring adjectives and the final visualization of the system. The end result is a visualization that compares the proportion of occurrence for each phrase for each brand. A simple version of this space is based on a list of predetermined adjectives. Through this visualization, the analysts obtained results such as the phrase 'ipod is overpriced' occurs 48 times as often as 'blackberry is overpriced'.

The complexity of the whole process revealed the power behind the visual management of information, otherwise difficult for the analysts to generate.

3.1.1 Impact report: Havas Media (Corporate Development: Gloria Andreu)

Before the existence of Impure, we had to contract several Opinion Mining agencies to extract online brand reputation. Impure has provided us an easy way to access, analyze and visualize such data without the need of such specialized agencies, reducing costs and increasing response time. Furthermore, Opinion Mining agencies are usually very topic specific, generating problems when comparing data from different types of markets.

Our company is now focusing on the development of a new market strategy based upon the usage of Impure to provide results from social data.

3.2 Study of correlations values of listed companies

Repsol's Department of Operational Models uses Impure as a system for analysis and visualization of distribution networks and other logistics data. One hypothesis we wanted to test is the relative impact on Repsol's behavior on the stock exchange of external market dynamics and internal dynamics, such as purchases, oil exploration results, changes in logistics or moves to enter new markets.

In this case, producer analysts generated a data flow with which to compare Repsol's stock market behavior with that of other companies (see Figure 3). This space allows the selection of different time intervals, in which analysts estimate the Pearson ratio between all pairs of a group of selected companies. The statistical correlation value was used as a weight in defining the different relationships between the tickers, allowing the creation of a network that enables these relationships to be seen in a force directed layout. In addition, the workspace includes a stacked plot which allows a temporal analysis of the data, which can be controlled by means of a slider that offers a dynamic view of the generation of a network, seeing the evolution of the Pearson correlations.

This workspace helped to determine when the company exhibited a behavior correlated with others in the same sector and when stock market behavior showed dynamic independence. Specifically, internal movements were identified that had the greatest impact on market independence and the delay in this impact could be measured.

3.2.1 Impact report: Repsol (Head of the Department of Operational Models: Jorge Aguirre)

The fact that Impure enables to represent and directly manipulate the data in a visual way has allowed us to communicate our analysis results in a better way to the non-expert board of directors, facilitating the dialogue with them.

Moreover, Impure helps in the process of building very specific dashboards to analyze the data. Such dashboards were previously ordered to third party companies that will be no longer needed.

From our experience, classic analysis tools present input and output processes, where the data transformation is obscured to the end users. Impure provides ways to create, follow and control the whole data flow, enlightening the analysis process.

4 Impact and future directions

Impure has created a big impact in both companies, modifying the way in which they analysts are exploring their data, and discussing the results with the board of directors. Both companies have also created new labs that will be specifically in charge of exploiting the capabilities of Impure.

According to our observations, the most valuable feature of the system is the possibility to use the visualizations as dynamic filters of the data, that can feed other modules, converting visualizations in part of the whole information flow rather than only acting as interpretation tools, as can be seen in the two workspaces presented. Moreover, the system takes care of the whole process that leads to the final visualizations, enabling the users to control in a visual way the data flow, facilitating the process of interiorizing the data.

Impure is being introduced in the analysis processes of several other major companies, which are working in a broad number of case studies. Users feedback will guide further development of the tool as well as drive new research questions.

5 Acknowledgements

We acknowledge the contribution of Repsol and Havas Media in providing feedback to our project.

References

[1] Tableau software, 2010. http://www.tableausoftware.com/.

[2] Yahoo pipes, 2010. http://pipes.yahoo.com/pipes/.

[3] J. Bresson, C. Agon, and G. Assayag. Visual lisp/clos programming in openmusic. *Higher Order Symbol. Comput.*, 22(1):81–111, 2009.

[4] M. Puckette. Pure data: another integrated computer music environment. In *in Proceedings, International Computer Music Conference*, pages 37–41, 1996.

[5] A. V. Riabov, E. Boillet, M. D. Feblowitz, Z. Liu, and A. Ranganathan. Wishful search: interactive composition of data mashups. In *WWW '08: Proceeding of the 17th international conference on World Wide Web*, pages 775–784, New York, NY, USA, 2008. ACM.

[6] F. B. Viegas, M. Wattenberg, F. van Ham, J. Kriss, and M. McKeon. Manyeyes: a site for visualization at internet scale. *IEEE Transactions on Visualization and Computer Graphics*, 13(6):1121–1128, 2007.

[1] http://words.bighugelabs.com/