ENHANCING BIG DATA EXPLORATION WITH FACETED BROWSING

Sonia Bergamaschi [1], Giovanni Simonini[1] and Song Zhu[1]

[1] Department of Engineering "Enzo Ferrari", Università di Modena e Reggio Emilia,
(e-mail: firstname.lastname@unimore.it)

KEYWORDS: Bayesian Network, Faceted Browsing, Big Data

With the modern information technologies, data availability is increasing at formidable speed giving raise to the Big Data challenge (Bergamaschi, 2014). As a matter of fact, Big Data analysis now drives every aspect of modern society, such as: manufacturing, retail, financial services, etc., (Labrinidis & Jagadish, 2012). In this scenario, we need to rethink advanced and efficient *human-computer-interaction* to be able to handling huge amount of data. In fact, one of the most valuable means to make sense of Big Data, to most people, is data visualization. As a matter of fact, data visualization may guide decision-making and become a powerful tool to convey information in all data analysis tasks. However, to be actually actionable, data visualization tools should allow the right amount of interactivity and to be easy to use, understandable, meaningful, and approachable.

In this article, we present a new approach to visualize and explore a huge amount of data. In particular, the novelty of our approach is to enhance the faceted browsing search in `Apache Solr`* (a widely used enterprise search platform) by exploiting Bayesian networks, supporting the user in the exploration of the data. We show how the proposed Bayesian suggestion algorithm (Cooper & Herskovits, 1991) be a key ingredient in a Big Data scenario, where a query can generate too many results that the user cannot handle. Our proposed solution aim to select best results, which together with the result-path, chosen by the user by means of multi-faceted querying and faceted navigation, can be a valuable support for both Big Data exploration and visualization.

In the fellowing, we introduce the *faceted browsing* technique, then we describe how it can be enhanced exploiting Bayesian networks.

Faceted Browsing. The faceted browsing (or faceted navigation) is a technique offered by many search engines for accessing information. It allows to

*http://lucene.apache.org/solr/resources.html

explore data applying dynamic filters in multiple steps: each time a filter is applied, the results are shown to the user, which can apply additional filters or modify existing ones. An example of a faceted panel is shown Figure 1. For each facet there are one or more values, called the facet values, used as filter for refining search query, interactively. Moreover, a facet counter may be associated to each value representing the number of records matching with this value. So, the faceted navigation allows the user to elaborate a query progressively, seeing the effect of each choice inside one facet on the available choices in other facets. From the user's perspective, faceted navigation eliminates the "dead ends" that may result from selecting unsatisfiable combinations of constraints among the facets. In fact, most combinations of facet values are unsatisfiable, because, the set of satisfiable combinations is typically a sparse subset of the set of all possible combinations. Furthermore, a search engine should not present to the user all facets and all facet values if the cardinality of the facet categories and their values is huge (condition often satisfied in the Big Data scenario) in fact, it would be a useless flood of information that cannot be reasonably handled by the user. Hence, the need of pruning techniques arise.

Our case study includes textual/semi-structured documents, where the number of facets reflects the number of ways a document can potentially be classified. In theory, there is no limit to the number of facets: there are infinitely many potential taxonomies to classify a document collection. In practice, of course, the number of facets is finite, but it may be quite large. There is also the issue of dependence among facets. For instance, if documents containing information about cities, states, and countries, we may devise either as three distinct facets or as a single hierarchical facet. On other hand, languages and nationalities are highly correlated and yet clearly distinct facets. At best, designing a faceted classification scheme with independent facets requires an extraordinary effort on the side of information architects; at worst, it is an impossible task as such a set of independent facets would not match the way users conceive the information space. Either way, we cannot require that facets be independent of one another.

Our work attempts to address these issues, by exploiting a probabilistic graphical model (i.e. Bayesian network) to capture facets dependencies and to determine the most valuable facets to be presented to users.

Improving Faceted Navigation with Bayesian Networks. Briefly, a Bayesian network (Nielsen & Jensen, 2009) is a compact representation of a probability distribution associated to a set of related variables. In our model, the variable modelled inside a Bayesian network are the attributes of a dataset, i.e.

the facets. Thus, Bayesian Networks can be exploited to infer the relationship among these facets. We develop our tool on top of `Apache Solr`, enhancing its faceted browsing interface, integrating facilities offered by `OpenMarkov`[†] to automatically learn a Bayesian network starting from data (e.g. from a `cvs` file). Thus, the tool shows how the search fields are dependent on each other in the form of graphs. The tool interface allows to give suggestion on the facets that she/he consider relevant on the graph, trying to find out other relevant features by using relationships among the attributes on the Bayesian Networks. In order to limit the number of items in each facet, the system calculates two groups of similar and dissimilar items, ranks them and returns a selection of the top *n* items to the user: The similar items help the user to define precise search, while the dissimilar ones stretch the range of search. Another key feature of our tool is the query recommendation, which can guide the user in formulating queries. In fact, when the user hovers over a facet in the `Solr` selection panel, the interface communicates to the tool the current query and the facet that the user intends to select. Then, it returns suggestions on the basis of how that choice would change probabilities of other facets, accordingly to the Bayesian network. So, the user is facilitated in his request as she/he will have a real-time feedback on the effects of his/her search task.

Testing the Tool. We modified the `Solr` front-end equipping it with a new faceted search interface and integrating it with a customized `OpenMarkov` API.

We tested our tool with a mushrooms dataset[‡], consisting of thousands of instances, each having 22 categorical attributes (e.g.: cap shape, cap surface, cap color, etc.). We employed the *Hill Climbing* algorithm to automatically learn the Bayesian network. The advantage of this algorithm is that it performs a heuristic to automatically search through the space of possible structures, using a metric that measures how well each structure can represent the probability distribution of the variables of the dataset (Cooper & Herskovits, 1991).

To conclude we present an example of the faceted browsing in our tool in the Figure 1. In the figure we show how the interface appears to a user hovering the mouse over *brown* of the *CapColor* facet. The pop-up shows which and how facets are affected, and what facets may be proposed to the user on the basis of the learned Bayesian network. In detail: the *unaltered* column contains the

[†] http://www.openmarkov.org/users.html

[‡] http://archive.ics.uci.edu/ml/datasets/Mushroom

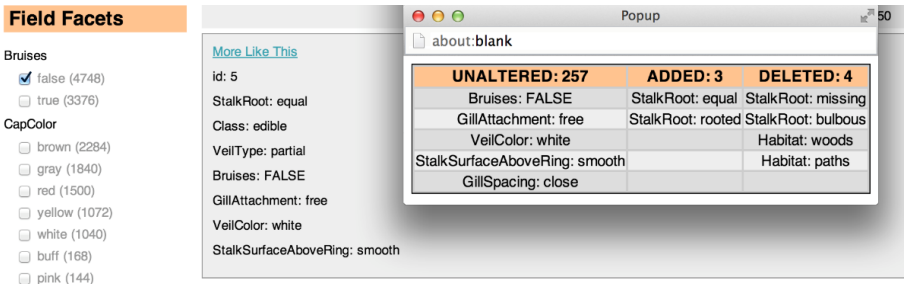

Figure 1. *Faceted browsing and query advisor on Solr.*

facet values present in the current selection and that would remain unaffected (i.e. they remains in the facet tab) after the application of the *brown* filter; the *added* column contains the facet values not present in the current selection and that would be added in the facet tab after the application of the *brown* filter; the *deleted* column contains the facet values present in the current selection would be deleted to the facet tab after the application of the *brown* filter.

References

BERGAMASCHI, SONIA. 2014. Big data analysis: Trends & challenges. *Pages 303–304 of: High Performance Computing & Simulation (HPCS), 2014 International Conference on*. IEEE.

COOPER, GREGORY F, & HERSKOVITS, EDWARD. 1991. A Bayesian method for constructing Bayesian belief networks from databases. *Pages 86–94 of: Proceedings of the Seventh conference on Uncertainty in Artificial Intelligence*. Morgan Kaufmann Publishers Inc.

LABRINIDIS, ALEXANDROS, & JAGADISH, HV. 2012. Challenges and opportunities with big data. *Proceedings of the VLDB Endowment*, **5**(12), 2032–2033.

NIELSEN, THOMAS DYHRE, & JENSEN, FINN VERNER. 2009. *Bayesian networks and decision graphs*. Springer Science & Business Media.

SIMON, HERBERT A. 1971. Designing organizations for an information-rich world.

TUNKELANG, DANIEL. 2009. Faceted search. *Synthesis lectures on information concepts, retrieval, and services*, **1**(1), 1–80.