SOURCERY: User Driven Multi-Criteria Source Selection

DOI:
10.1145/3269206.3269209

Document Version
Accepted author manuscript

Published in:
Proceedings of the 27th ACM International Conference on Information and Knowledge Management

SOURCERY: User Driven Multi-Criteria Source Selection

Edward Abel, John A. Keane, Norman W. Paton, Alvaro A.A. Fernandes, Martin Koehler, Nikolaos Konstantinou, Nurzety A. Azuan, Suzanne M. Embury
School of Computer Science, University of Manchester, UK

ABSTRACT

Data scientists are usually interested in a subset of sources with properties that are most aligned to intended data use. The SOURCERY system supports interactive multi-criteria user-driven source selection. SOURCERY allows a user to identify criteria they consider of importance and indicate their relative importance, and seeks a source selection result aligned to the user-supplied criteria preferences. The user is given an overview of the properties of the sources that are selected along with visual analyses contextualizing the result in relation to what is theoretically possible and what is possible given the set of available sources. The system also enables a user to interactively perform iterative fine-tuning to explore how changes to preferences may impact results.

CCS CONCEPTS

• **Information systems → Information retrieval**; • **Applied computing → Decision analysis**;

KEYWORDS

Source selection, Multi-criteria decision analysis, optimization

ACM Reference Format:
Edward Abel, John A. Keane, Norman W. Paton, Alvaro A.A. Fernandes, Martin Koehler, Nikolaos Konstantinou, Nurzety A. Azuan, Suzanne M. Embury. 2018. SOURCERY: User Driven Multi-Criteria Source Selection. In *The 27th ACM International Conference on Information and Knowledge Management (CIKM '18), October 22–26, 2018, Torino, Italy.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3269206.3269209

1 INTRODUCTION

There are increasingly many data sources, from organizations producing numerous internal sources [6] and technological developments such as web data extraction [4]. As a result, data scientists are often only interested in a subset of these available sources. The most important criteria for informing source selection, and their relative importance, are likely to be both *user*-specific and *use*-specific.

The SOURCERY system supports interactive multi-criteria source selection, allowing a user to identify the set of relevant criteria, and to define their relative importance. An example in Figure 1 shows user preferences regarding a set of criteria relating to a set of available data sources from the real-estate domain. Here, each drop down box selection denotes user preference between a pair of criteria via descriptive terms such as `slightly more important`, `strongly more important`, etc., that denote which of the pair is most important and by how much. The first combo box involves *postcode quality* and *tuple completeness*; it denotes that *tuple completeness* is `strongly less important` than *postcode quality*. SOURCERY uses these preferences to determine, from the set of available sources, which should be selected to produce a dataset best aligned with user preferences. The data scientist can then be presented with information about, and analysis of, the sources to be selected.

Although user preferences drive source selection, the result is likely to involve certain trade-offs in the multi-dimensional space of sources characterized by the criteria values. In this context, the results of source selection may not meet user expectations, and what is actually possible via the available data sources may be unclear. To support the user to explore this space, further analysis contextualizes the result, allowing the user to visualize how the result compares with both what is theoretically possible and what is possible given available data sources. The system allows a user to interactively fine tune the result to explore, for example, the impact on results of a slight change in preferences. Previous results are retained so that the user can assess this impact; when the user is happy with the resulting source selection, the selection can be realized and the resulting dataset presented to the user.

2 TECHNICAL OVERVIEW

For a set of sources, given a set of criteria, typically some sources will contain higher quality data according to some criteria but lower quality according to other criteria. The SOURCERY system tackles source selection as a Multi-Criteria Decision Analysis (MCDA) problem [8], enabling users to formalize their preferences regarding a set of criteria by specifying relative importance using Pairwise Comparison (PC). In such a multi-criteria scenario, a range of possible trade-off solutions exist. Sources are selected from the objective space to find a trade-off solution aligned with user preferences. This contrasts with other results on source selection, which map the inherently multi-criteria space onto a single criterion for optimization, e.g. [3] and [12], or explore the multi-criteria space without considering user preferences and how to elicit them [11], instead calculating multiple trade-off solutions by sampling different weighting configurations between a set of criteria [10].

2.1 Preference Elicitation

For a set of C criteria, PC allows a user to consider one pair of criteria at a time, and to define preference, and strength of preference, between the pair. This induces a separation of concerns:

(1) helping to achieve an accurate reflection of user preferences, compared to other elicitation techniques, such as direct numeric elicitation [9];
(2) aiding a user to both form and clarify his/her preferences [7] in an intuitive and user-friendly manner [5].

CIKM '18, October 22–26, 2018, Torino, Italy

ACM ISBN 978-1-4503-6014-2/18/10…$15.00
https://doi.org/10.1145/3269206.3269209

From a completed set of PC, one for each pair, a vector of criteria weights can be derived. Within SOURCERY, criteria weights are calculated using the Geometric Mean (GM) prioritization method [2] which, first compiles the set of PC into a PC-matrix (containing the set of PC along with their reciprocal and self comparisons) and then calculates weights, from this matrix, as the product of each row raised to the inverse power of C. These weights are then normalized to sum to 1; see [2].

2.2 MinSum Optimization

Source selection, by definition, requires a means to be able to discriminate between those sources which are selected and those which are not. In SOURCERY this is defined through a user-specified threshold of the minimum resulting dataset size (in terms of number of tuples). Given a set of possible sources S, a minimum resulting dataset size B, and a set of criteria C and weights $W = [w_1, w_2, ..., w_c]$ modeling user preferences between criteria, SOURCERY identifies a set of sources most aligned with user preferences, via the MinSum Optimization strategy [1], from which tuple selection can be performed. MinSum considers user preferences in determining a trade-off solution that has minimum overall weighted deviation from the set of criteria's ideal values. First MinSum finds the ideal solution Z^* for each criterion separately via single objective optimization; each represents an optimal solution for a single criterion, satisfying constraints of the requested result size threshold and the size of each source. Similarly, the negative ideal solution Z^{**} for each criterion is found in the same way; each represents the worst possible solution for a criterion. MinSum uses this information, along with user criteria weights, to find a trade-off solution, that minimizes the sum of the set of weighted criteria deviations. Each weighted criterion deviation is a measure of the compromise within a solution in relation to each criterion's ideal value, weight adjusted to reflect user preferences.

The weighted deviation for criterion j, D_j, is calculated via:

$$D_j = \frac{w_j(Z_j^* - \sum_{i=1}^{|S|} q_i \cdot c_{ij})}{(Z_j^* - Z_j^{**})} \quad (1)$$

where Z_j^* is the ideal solution for criterion j , Z_j^{**} the negative ideal solution for criterion j, W_j the weight of criterion j, c_{ji} the quality measure of criterion j for source i, and q_i the amount of data in source i (its total size).

A criterion's weighted deviation therefore is a weight adjusted measure of how close a solution is to the ideal value for that criterion, and the higher the criterion's weight the larger its deviation will appear in comparison to other criterion's deviation. MinSum seeks to minimize the overall sum of the set of weighted criteria deviations, in this way seeking a solution where the trade-offs are aligned with user preferences. Therefore, MinSum optimises:

$$min \sum_{j=1}^{|C|} D_j \quad (2)$$

where D_j represents the weighted deviation value for criterion j, see (1). The model is solved subject to the requested result size threshold, the size of each source and with an additional set of constraints to determine the weighted deviation for each criterion. From the MinSum optimization a set of sources are identified, which constitute the source selection result.

3 THE SYSTEM

We introduce the SOURCERY system, which facilitates, within a browser, interactive multi-criteria user-driven source selection. A real-world dataset consisting of web-scraped data from the real-estate domain, extracted via the DIADEM system [4], is used as a case study. The dataset contains 36,818 tuples, each relating to a single Property, from 137 UK real-estate sites. For each Property

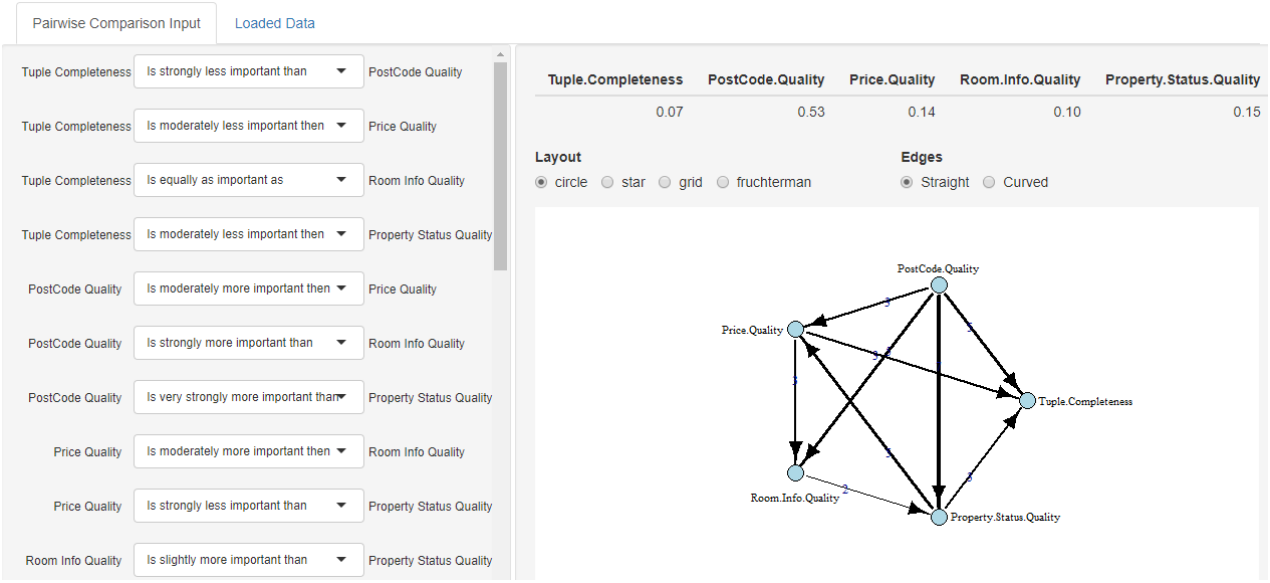

Figure 1: PC elicitation and analysis

Figure 2: Selected sources and their properties analysis

DIADEM aims to obtain its location, its room composition and other particulars. The quality of such data varies from source to source, due to the amount of information that is extractable as well as the success of the extraction process. The quality of the extracted data from each source can be assessed with respect to various criteria, such as *tuple completeness* and *postcode quality*[1], for which we obtain a measure of each source's average quality. Such measures for each source for each criterion can be estimated from sampling. Therefore, for a given dataset we have a set of possible criteria, which can be made up of both domain-agnostic criteria, such as *tuple completeness*, and more domain-specific ones, such as *postcode quality*. For additional or new criteria, given a defined function for the criterion, it could be utilized, for example, over a sample from each source to determine values for the sources for the new criterion. Then the new criterion could be added to the list of available criteria for the dataset. Therefore, for a given dataset, along with the sources themselves, such a meta-data file contains information pertaining to a set of, potentially all, criteria.

3.1 Input and Preference Elicitation

The user selects a dataset of available sources, such as the real-estate dataset, and SOURCERY loads the data pertaining to the criteria for the set of available sources of the dataset. SOURCERY allows a user to explore this information to see, for example, the spread of values over the sources for different criteria. From the set of possible criteria for the selected dataset, the user selects the subset of relevant criteria. With the subset of chosen criteria, the user can perform PC, via a set of combo boxes, as shown on the left side of Figure 1. With every PC modification by the user, the system updates the set of criteria weights calculated from the comparisons (as shown in the top right of Figure 1) and computes the source selection result. Dynamic computing of the source selection result can be toggled to be manual for a user wishing to define multiple modifications before updating. The associated Directed Acyclic Graph (DAG), as shown on the right of Figure 1, is also updated with every PC modification. The DAG gives a visual representation of the set of comparisons, with criteria as nodes and comparisons and their strengths as labeled edges.

3.2 Analysis of Source Selection Results

For a source selection result the user is given a visual overview of selected sources and their size, as shown in Figure 2. Each criteria value for each selected source, normalized with respect to the range of each criterion within the set of available sources, is shown along with each source's overall criteria values sum by which selected sources are ranked. The selected source values can be further analysed by toggling if they are shown weight-adjusted, with respect to the criteria weights, and the selected sources can be ranked by their overall weighted criteria values sum. Further analysis visually contextualizes the result, broken down by each criterion, regarding how the result compares in relation to what is theoretically possible for each criterion, and what is possible given the available sources (and the size of result threshold), as shown in Figure 3.

Analysis of such figures enables a user a deeper understanding of their datasets and the trade-offs involved in his/her source selection solution. For example, from Figure 2 - that shows the resulting source selection result from the preferences expressed in Figure 1 - the user can ascertain that *price quality* is generally very high across all selected sources, whilst *postcode quality*, whilst of high quality in some of the selected sources, is nothing in many sources, due to missing data. Further, Figure 3 shows analysis of the source selection result derived from the preferences expressed in Figure 1 - showing the user that although postcode quality in the result is low compared to the theoretical highest, in terms of what is possible given available sources, it is very close to the highest possible, in keeping with the high weight assigned to it by the user. Such analysis and conceptualization of the trade-offs of a result aid validation and traceability of a source selection result and help a user to ascertain what is possible and thus realistic. If the user is happy with the source selection result, then it can be realized and the resulting selected tuple-set presented to the user. Alternatively, the result can be fine-tuned.

[1]See [1] for full definitions and calculations of these and further criteria for the dataset.

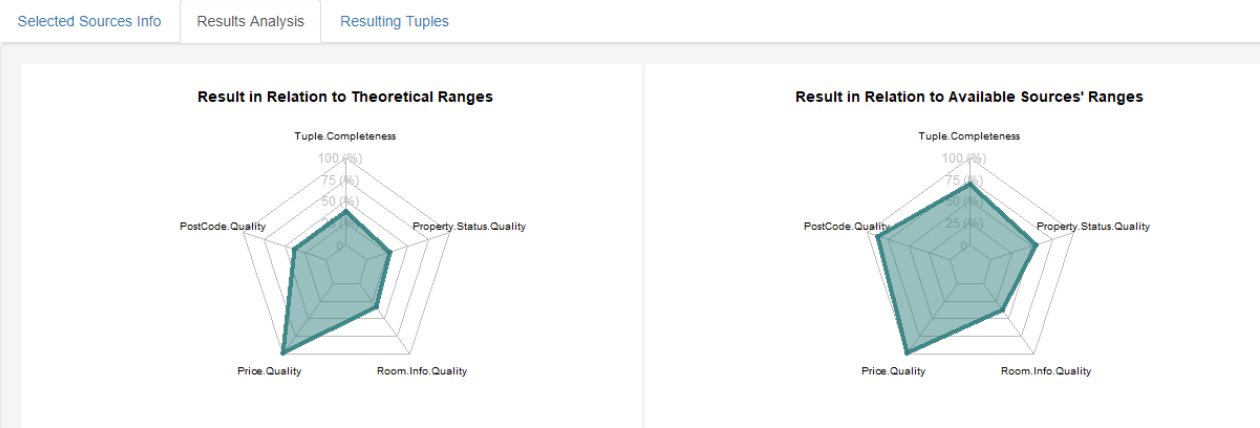

Figure 3: Analysis of what is theoretically possible (left), and what is possible given set of available data sources (right)

3.3 Result Fine-tuning

The system allows a user to fine tune the source selection result by exploring how changes to preferences impact the result. For example, consider a user seeking high *tuple completeness* and high *postcode quality* who, after setting initial preferences and analysing Figure 3, decides that *tuple completeness* in relation to what is possible is insufficient; preferences can then be changed to increase the relative importance of *tuple completeness*. The impact upon the result can be viewed in updated spider plots, and the user can determine if *tuple completeness* in relation to what is possible is now sufficient, and that any possible loss of quality in other criteria to obtain higher *tuple completeness*, due to trade-offs between the criteria, is tolerable. Similairy, the user can analyse how such changes impact upon the sources that are chosen and their properties in the selected sources' information plot. To help a user perform such analysis, previous results are retained to enable the user to easily assess how a change impacts results. The system supports interactive response times to obtain a selection result involving thousands of sources and several criteria.[2]

4 DEMONSTRATION

SOURCERY will be demonstrated as a web application, using the real-world web-scraped real-estate dataset outlined in Section 3. The demonstration will allow a user to select criteria, and to define their relative preferences via PC (Figure 1). After each PC alteration, or alteration of the result size threshold parameter, the selected sources analysis and contextualization analysis are provided in real-time. The user can explore the result via a visual overview of sources selected and their properties (Figure 2), and a contextualization analysis showing for each criterion how the result compares in terms of both what is theoretically possible and what is possible given the set of available sources (Figure 3). Such analyses convey to a user the characteristics of the available sources and how they impact what is possible within a result; iterative fine tuning facilitates refinement of a solution until a satisfactory result is obtained. Finally, the source selection result will be realized and the resulting tuples presented.

[2] A detailed performance analysis of the optimization algorithm is given in [1]

5 CONCLUSIONS

The SOURCERY system supports interactive multi-criteria user-driven source selection to support data scientists, who are often only interested in a subset of available sources, to find data with properties that reflect intended use. This demonstration illustrates how PC can aid in eliciting a user's preferences between a set of criteria, and how analyses over the results can provide an interactive, exploratory and well informed source selection activity.

Acknowledgment This work is supported by the VADA Programme Grant of the UK Engineering and Physical Sciences Research Council, grant number EP/M025268/1.

REFERENCES

[1] Edward Abel, John Keane, Norman W. Paton, Alvaro A.A. Fernandes, Martin Koehler, Nikolaos Konstantinou, Julio Cesar Cortes Rios, Nurzety A. Azuan, and Suzanne M. Embury. 2018. User driven multi-criteria source selection. *Information Sciences* 430-431 (2018), 179–199.
[2] G B Crawford. 1987. The geometric mean procedure for estimating the scale of a judgement matrix. *Mathematical Modelling* 9, 3-5 (1987), 327–334.
[3] Xin Luna Dong, Barna Saha, and Divesh Srivastava. 2012. Less is more. *Proceedings of the VLDB Endowment* 6, 2 (2012), 37–48.
[4] Tim Furche, Georg Gottlob, Giovanni Grasso, Xiaonan Guo, Giorgio Orsi, Christian Schallhart, and Cheng Wang. 2014. DIADEM: Thousands of Websites to a Single Database. *PVLDB* 7, 14 (2014).
[5] Ixent Galpin, Edward Abel, and Norman W Paton. 2018. Source Selection Languages: A Usability Evaluation. In *Proceedings of the Workshop on Human-In-the-Loop Data Analytics*. ACM, 1–6.
[6] Alon Halevy, Flip Korn, Natalya F. Noy, Christopher Olston, Neoklis Polyzotis, Sudip Roy, and Steven Euijong Whang. 2016. Goods: Organizing Google's Datasets. In *SIGMOD*. ACM Press, New York, New York, USA, 795–806.
[7] A Ishizaka, D Balkenborg, and T Kaplan. 2011. Does AHP help us make a choice? An experimental evaluation. *JORS* 62, 10 (2011), 1801–1812.
[8] A Ishizaka and Phillipe Nemery. 2013. *Multicriteria decision analysis: methods and software*. Wiley.
[9] Ido Millet. 1997. The Effectiveness of Alternative Preference Elicitation Methods in the Analytic Hierarchy Process. *Journal of Multi–Criteria Decision Analysis* 6, 1 (1997), 41–51.
[10] Theodoros Rekatsinas, Amol Deshpande, Xin Luna Dong, Lise Getoor, and Divesh Srivastava. 2016. SourceSight: Enabling Effective Source Selection. In *ACM SIGMOD*. 2157–2160.
[11] Theodoros Rekatsinas and Lise Getoor. 2015. Finding Quality in Quantity : The Challenge of Discovering Valuable Sources for Integration. *CIDR* (2015).
[12] Mariam Salloum, Xin Luna Dong, Divesh Srivastava, and Vassilis J. Tsotras. 2014. Online Ordering of Overlapping Data Sources. In *VLDB 2014*. 133–144.