From Self-data to Self-preferences: Towards Preference Elicitation in Personal Information Management Systems

Tristan Allard[1(✉)], Tassadit Bouadi[1(✉)], Joris Duguépéroux[1,2], and Virginie Sans[1]

[1] Univ. Rennes 1/IRISA, Rennes, France
{tristan.allard,tassadit.bouadi,joris.dugueperoux, virginie.sans}@irisa.fr
[2] ENS Rennes, Rennes, France
joris.dugueperoux@ens-rennes.fr

Abstract. Ever-increasing quantities of personal data are generated by individuals, knowingly or unconsciously, actively or passively (*e.g.,* bank transactions, geolocations, posts on web forums, physiological measures captured by wearable sensors). Most of the time, this wealth of information is stored, managed, and valorized in isolated systems owned by private companies or organizations. Personal information management systems (PIMS) propose a groundbreaking counterpoint to this trend. They essentially aim at providing to any interested individual the technical means to re-collect, manage, integrate, and valorize his/her own data through a dedicated system that he/she owns and controls. In this vision paper, we consider personal preferences as first-class citizens data structures. We define and motivate the *threefold preference elicitation problem in PIMS* - elicitation from local personal data, elicitation from group preferences, and elicitation from user interactions. We also identify hard and diverse challenges to tackle (*e.g.,* small data, context acquisition, small-scale recommendation, low computing resources, data privacy) and propose promising research directions. Overall, we hope that this paper uncovers an exciting and fruitful research track.

Keywords: Preferences · Privacy · Self-data
Personal information management system

1 Introduction

An ever-increasing quantity and diversity of personal data feeds the database systems of various companies (*e.g.,* emails, shopping baskets, news, geolocations, physiological measures, electrical consumption, movies, social networks, posts on forums, professional resumes). Although individuals often benefit indirectly from this large-scale systematic capture of their data (*e.g.,* free access to services), the use value they get from it remains strongly limited by its fragmentation

R. Guidotti et al. (Eds.): PAP 2017, LNCS 10708, pp. 10–16, 2017.
https://doi.org/10.1007/978-3-319-71970-2_2

in non-cooperative *data silos* and by the usage allowed and supported by each silo [1,5]. Personal information management systems (*PIMS* for short) aim at giving to individuals technical means to re-collect, integrate, manage, and use their data (or at least a part of it) in a central location *under their control* (*e.g.,* a personal computer, a rented virtual machine). The expected uses of a PIMS are those of a full-fledged data-centric personal assistant, including for example panoramic integrated personal data visualizations (*e.g.,* health data from health centres and physiological measures from wearables), vendor relationship management and privacy-preserving recommendations (*e.g.,* comparing offers from electricity providers given one's detailed electrical consumption), automatic completion of online profiles (*e.g.,* privacy preferences on a social network)[1].

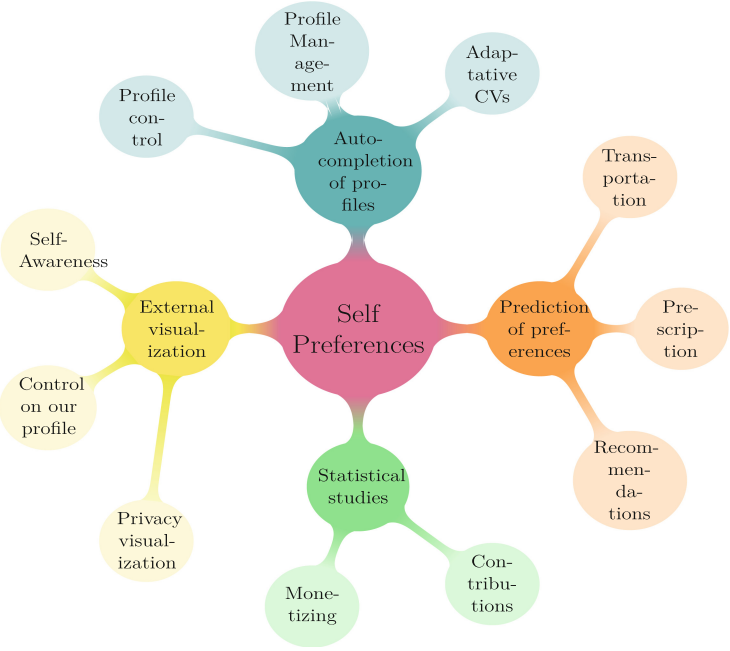

Fig. 1. A non-exhaustive list of uses of self-preferences

We are moreover experiencing today a strong push towards the widespread adoption of PIMS: various commercial industrial-strength PIMS exist today (*e.g.,* Cozy[2], Hub of All Things[3]), modern laws in favor of data portability are passed (*e.g.,* article 20 of the new European General Data Protection Regulation[4], article 48 of the new French data protection bill[5]), institutions

[1] See, *e.g.,* https://tinyurl.com/mesinfosValue for a large variety of use-cases.
[2] https://cozy.io/en/.
[3] http://hubofallthings.com/.
[4] https://tinyurl.com/euDataPort.
[5] https://tinyurl.com/frDataPort.

launch initiatives and express opinions in favor of PIMS (*e.g.,* the European Data Protection Supervisor opinion about PIMS[6], the US MyData initiatives[7], the UK MiData initiative[8]), and the research field is active (*e.g.,* [1,2,5,6]).

However the promesses that PIMS carry on will be strongly hampered if they fail in eliciting *personal preferences* from the wealth of personal data they store, simply because they crucially rely on accurately modeling, reasoning, and using personal preferences. Consider for example an automatic profile completion application based on preferences (see Fig. 1 for other examples of uses). It is often time-consuming and error-prone to fill in the forms about, *e.g.,* privacy preferences. A detailed automatic completion application based on preferences would for example precisely tune the privacy preferences of social networks depending on their application domains (*e.g.,* a friendship-centric network would be parameterized differently than a professional social network).

In this vision paper, we advocate for considering personal preferences as first-class citizens in PIMS. We call them *self-preferences*, characterize the PIMS computing environment, and identify key challenges to be solved for eliciting self-preferences from the personal data stored in a PIMS. These challenges fall in three categories: local elicitation (based on the data that is stored locally), global elicitation (or group elicitation, based on the data of similar individuals), and interactive elicitation (based on the individual's feedback). They involve various fields of research such as, *e.g.,* preference elicitation, small data, data integration, data privacy, distributed computing, data visualization.

This paper is organized as follows. Section 2 defines more precisely the problem of self-preferences elicitation in PIMS. Section 3 identifies the main related key challenges to be tackled. Finally, Sect. 4 concludes.

2 Problem Statement

2.1 Personal Information Management Systems

A *Personal Information Management Systems* (or *PIMS* for short) is basically a suite of software in charge of collecting, centralizing, and managing the personal data related to an individual. In this work, we focus on PIMS that are hosted and run on a server *controlled by the individual* [1] (*e.g.,* owned or rented). A PIMS may be executed in a large variety of settings: a personal device possessed by the individual (*i.e.,* self-hosting model - Cozy for example), a virtual machine rent by the individual to a cloud provider (*i.e.,* platform-as-a-service model - Hub of all things for example). Despite such heterogeneity, we believe that most PIMS can be characterized as follows:

Resources. Whatever the PIMS execution environment, it is dedicated to serve a single individual, or at most a few closely related individuals (*e.g.,* the members of a family). Therefore, the local computing resources—CPU and RAM—are

[6] https://tinyurl.com/edpsOnPims.
[7] https://tinyurl.com/usMyData.
[8] https://tinyurl.com/ukMiData.

scarce. Typical worst-case resources are on the order of those provided by a low-cost personal device (*e.g.,* a raspberry-pi—around 1 GB RAM and 1.2 GHz CPU), although on average they are probably similar to current commodity hardware. More resources may be available when the execution environment is a virtual machine in the cloud, but they remain limited to the personal space of the individual.

Threat Models. First, a PIMS is trusted by its owner. It is actually a prerequirement because it stores the personal data together with the credentials required for fetching data from the web (couples of login/password). Second, a PIMS is not necessarily trusted by other individuals. For example, it may be hosted by a controversial cloud provider, or self-hosted and possibly ill-protected. When several PIMS collaborate for performing a global computation (*e.g.,* statistics over a population), the personal data involved in the computation need to be protected from dubious PIMS. Typical attack models include the *honest-but-curious* model (observes any information that leaks during the protocol) and the *malicious* model (may additionally deviate from the protocol - *e.g.,* by forging or tampering messages). Resistance to *collusions* of PIMS must also be considered.

2.2 Preference Model

Preferences express comparisons on a set X of items, choices or alternatives. Preferences could be expressed in different forms. We can express preferences in a quantitative way, by expressing degrees of interest in terms of scores (*e.g., "My preference for red wine is 0.5 and for white wine is 0.3"*), or in a qualitative way, by pairwise comparisons or other AI preference formalisms (*e.g., "I prefer red wine to white wine"*). The qualitative approach is more general than the quantitative one.

More formally, a preference relation (*i.e., preference order*) $\succeq \subseteq X \times X$ is a *reflexive, antisymmetric and transitive binary relation*, where $x \succeq y$ means *"x is at least as good as y"*. Different types of preference orders can be defined, depending on the properties they satisfy (symmetry, transitivity, etc.).

Furthermore, preferences are of various kinds and take different forms (*e.g., fuzzy preferences, conditional preferences, bipolar preferences, etc.*). Several preference modeling formalisms have been proposed in the literature, each of which allows to express and model some properties of preferences. Thus, by comparing the expressive power of these formalisms, one can choose the adequate formalism that accurately describes and captures the rich cognitive model of an individual.

In real-life applications, preferences fully depend on the decision context. We believe that context-aware models such as, *e.g.,* the conditional preference network formalism (CP-net) [4] is especially adequate for self-preferences. This formalism provides an intuitive way to specify conditional or contextual statements of the form *"If context or condition C is true, then preference P is true"* (*e.g., "If the main course is meat, I prefer red wine to white wine"*).

2.3 Elicitation of Self-preferences: A Threefold Problem

The general purpose of this work is to elicit and maintain the self-preferences of an individual, in a PIMS context, for letting him/her use, manage, and visualize them. The problem is actually threefold, depending on the actual source of information used for the self-preference elicitation (we do not pretend to have listed all the potential information sources for the self-preference elicitation problem).

Problem 1: Local Elicitation. The primary source of information is the personal data stored in the PIMS. Problem 1 is thus the local elicitation of self-preferences. It can be phrased as follows: *which model(s) would be adequate for self-preferences and how to elicit them on commodity hardware based on heterogeneous and small data?*

Problem 2: Global Elicitation. The secondary source of information is the self-preferences of groups of *similar* individuals. Indeed, PIMS are usually connected to the Internet (even if they may get disconnected arbitrarily) and can participate in distributed algorithms provided that the personal data and preferences they store remain adequately protected. The preferences of similar individuals may thus enrich local self-preferences. Problem 2 is thus the global, group-based, elicitation of self-preferences: *How to form meaningful group preferences shared by a significant number of individuals—in a* privacy-preserving manner *and from commodity hardware devices—and enrich local self-preferences accordingly?*

Problem 3: Manual Elicitation. Finally, the third available source of information is the individual him/her-self. Moreover, in a real-life context, self-preferences are not static (*e.g.,* change of beliefs, new self-preferences learned, new data sources feeding the PIMS). As a result, problem 3 is the refinement and revision by the individual of the self-preferences already stored in his/her PIMS (*e.g.,* elicited locally or globally). *How to guide a non-expert individual for letting him/her interactively refine or revise his/her self-preferences and perform the updates efficiently on commodity hardware?*

3 Challenges

We identify below key challenges to be addressed for solving each of the three problems stated in Sect. 2.

Challenge 1: Local is Small. The idea of accurately reflecting and expressing individuals' preferences is a common and known AI issue (see, *e.g.,* preference mining [8] and preference learning [7] techniques). Consequently, previous works have coped with this problem, leading to a sophistication of the preference models and thus to more tedious and resource-consuming elicitation methods. Proposing effective preference elicitation techniques becomes more and more challenging when facing *small and heterogeneous data* and having *low computing resources*. First, we have to *integrate and reason* about data from different sources, and this in a consistent manner. We think that providing the system

with *context metadata* will ease the integration process. Second, the *small scale of local data* is a challenge to overcome for accurately learning even the simplest models. Lastly, we have to *develop optimal and efficient local elicitation algorithms* to address the low computing resources of a typical PIMS.

Challenge 2: Global is Threatening. In order to benefit from the self-preferences of similar individuals for enriching the local self-preferences, three steps are necessary: (1) groups of similar individuals must be formed, (2) global group preferences must be aggregated from local self-preferences, and (3) *some* group preferences must be selected for enriching local self-preferences. Performing these steps (possibly combined together) in a distributed and privacy-preserving manner is a difficult challenge. Several previous works have already tackled problems similar to step 1 (*e.g.,* privacy-preserving distributed clustering, k-nearest neighbors)—although usually the underlying data is not preferences or the underlying model is not as sophisticated as elaborate preference models. We believe that the main difficulties lie in step 2 and step 3. Indeed, aggregating together a set of preferences (see, *e.g.,* [3]) may require to perform operations not supported efficiently by encryption schemes (*e.g.,* comparison operators for implementing `MIN` aggregate, or threshold computations). Moreover, the low computing resources of PIMS may not be able to cope with encryption-intensive algorithms, calling for alternative protection-by-perturbation strategies (*e.g.,* differential privacy) and consequently the design of specific perturbation mechanisms for preferences and specific privacy/utility tradeoffs. Finally, the challenge gets even harder when considering that preference models may be heterogeneous across PIMS.

Challenge 3: Manual is Tedious. With the possibility of interactive revision and refinement, many challenges arise. Revising and refining contextual preferences may require to merge (1) preferences of the same context, (2) preferences with context hierarchically associated, (3) preferences with different contexts having similar ontologies. This leads to design robust techniques for the acquisition of the proposed preference model. Furthermore, these techniques should fit limited-resource systems. The interactivity aspect will also benefit from comprehensive preference-human interfaces, that should provide mechanisms to express rich preference models while keeping user effort to an acceptable level.

4 Conclusion

In this vision paper, we discuss the value of self-preferences for their owners, foreshadow an illustrative but not exhaustive list of the potential self-preferences uses, define the *threefold self-preference elicitation problem in PIMS*, and identify key research challenges that need to be tackled in order to help individuals to express, use, and manage their self-preferences. We hope this paper highlights promising research avenues in the field of self-data management.

References

1. Abiteboul, S., André, B., Kaplan, D.: Managing your digital life. Commun. ACM **58**(5), 32–35 (2015)
2. Abiteboul, S., Arenas, M., Barceló, P., Bienvenu, M., Calvanese, D., David, C., Hull, R., Hüllermeier, E., Kimelfeld, B., Libkin, L., Martens, W., Milo, T., Murlak, F., Neven, F., Ortiz, M., Schwentick, T., Stoyanovich, J., Su, J., Suciu, D., Vianu, V., Yi, K.: Research directions for principles of data management (abridged). SIGMOD Rec. **45**(4), 5–17 (2017)
3. Basu Roy, S., Lakshmanan, L.V., Liu, R.: From group recommendations to group formation. In: Proceedings of SIGMOD 2015, pp. 1603–1616 (2015)
4. Boutilier, C., Brafman, R.I., Domshlak, C., Hoos, H.H., Poole, D.: CP-nets: a tool for representing and reasoning with conditional ceteris paribus preference statements. J. Artif. Intell. Res. **21**, 135–191 (2004)
5. de Montjoye, Y., Wang, S.S., Pentland, A.: On the trusted use of large-scale personal data. IEEE Data Eng. Bull. **35**(4), 5–8 (2012)
6. Estrin, D.: Small data, where N = Me. Commun. ACM **57**(4), 32–34 (2014)
7. Fürnkranz, J., Hüllermeier, E.: Preference learning: an introduction. In: Fürnkranz, J., Hüllermeier, E. (eds.) Preference learning, pp. 1–17. Springer, Heidelberg (2010). https://doi.org/10.1007/978-3-642-14125-6_1
8. Holland, S., Ester, M., Kießling, W.: Preference mining: a novel approach on mining user preferences for personalized applications. In: Lavrač, N., Gamberger, D., Todorovski, L., Blockeel, H. (eds.) PKDD 2003. LNCS (LNAI), vol. 2838, pp. 204–216. Springer, Heidelberg (2003). https://doi.org/10.1007/978-3-540-39804-2_20