Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 52 (2015) 1040 – 1046

Procedia Computer Science

International Workshop on Big Data and Data Mining Challenges on IoT and Pervasive Systems (BigD2M 2015)

Refinement Strategies for Correlating Context and User Behavior in Pervasive Information Systems

Ali Jaffal*, Bénédicte Le Grand, Manuele Kirsch-Pinheiro

Centre de Recherche en Informatique / Université Paris 1 – Panthéon Sorbonne, 90 Rue Tolbiac, Paris 75013, France

Abstract

Large amounts of traces can be collected by Pervasive Information Systems, reflecting user's actions and the context in which these actions have been performed (location, date, time, network connection, etc.). This article proposes refinement strategies with different frequency measurements on contextual elements in order to better analyze the impact of these elements on the user's behavior. These strategies are based on data mining and Formal Concept Analysis and used to refine input data in order to identify the context elements that have a strong impact on user behaviors. We go further on context analysis by cognizing FCA with semantic distance measures calculated based on a context ontology. The proposed context analysis is further on evaluated in experiments with real data. The novelties of this work lies on these refinement strategies which can lead to a better understanding of context impact. Such understanding represents an important step towards personalization and recommendation features.

Peer-review under responsibility of the Conference Program Chairs

Keywords: Ubiquitous Computing; Context-aware computing; data mining; Formal Concept Analysis;

1. Introduction

Pervasive Information Systems (PIS) constitute an emerging class of Information Systems, in which information technology is gradually embedded in the physical environment[1]. Huge volumes of data can be generated from these systems by observing user interactions and the environment in which they take place. Such data is extremely valuable for systems designers and service providers, who can enhance their systems and services according to

* Corresponding author. Tel.: +33-144-078-604; fax: +33-144-078-954.
E-mail address: ali.jaffal@malix.univ-paris1.fr

Peer-review under responsibility of the Conference Program Chairs
doi:10.1016/j.procs.2015.05.103

user's context (e.g., location, date, time, past and current activities). Context can be defined as any information that can be used to characterize the situation of an entity (a person, place, or object) that is considered relevant to the interaction between a user and an application[2]. In this article, we propose to investigate the relative impact of various context elements on the user's activities. Context elements may have an influence on the user's choices. By understanding this influence, it is possible to propose better personalization and recommendation features and to improve adaptation mechanisms. However, identifying *relevant* context elements (for a given user) from the highly heterogeneous collected context data is a challenge not only because of data heterogeneity, but also because this relevance is personal and may vary significantly from user to user. Context analysis methods are then necessary in order to establish context relevance for a given user. In earlier work[3], we have used Formal Concept Analysis (FCA) to this end and shown its interest to cluster context elements with regard to associated user activities (and vice versa) into overlapping classes. However, FCA's strength becomes a weakness in the context of Big Data: all input data are considered as equally important, and the interpretation of results is complex. The contribution of this paper consists in proposing a methodology to reflect in FCA results the relative importance of context elements on user activities, with regard to their frequencies. We propose and compare two frequency measures on context elements, which we use to refine the data given as input to FCA algorithms. We propose two refinement strategies for each frequency measure, and we experiment them on real data collected from tablet users.

The paper is structured as follows: Section 2 presents related works and introduces FCA. Section 3 introduces the proposed methodology, while Section 4 presents its evaluation. Section 5 concludes the paper.

2. Related works

Context-awareness can be defined as the ability a system has to adapt its operations to the current context[4]. It represents an important requirement for Pervasive Information Systems (PIS) that must adapt their behavior in order to offer user relevant services and information. According to Greenberg[5], several factors contribute to the context and its relevance highly depends on the particular situation. Analyzing the impact of context information on the user's actions becomes a key aspect for successful PIS.

Data mining is sought to satisfy the growing demand for intelligence on pervasive environments. It is promoted as the research for relevant information that supports decision and prediction[6,7]. Several works such as [8] are considering data mining techniques for analyzing context data. These techniques can play an important role in understanding and discovering the intrinsic relationships between data. Among them, Formal Concept Analysis (FCA)[9,10] represents an interesting method of conceptual clustering. It provides an approach to knowledge organization. It helps finding a natural data structure, while associating actions to observed context elements.

The input of the FCA algorithms is a binary relation - called *formal context* - between a set of *objects* and a set of *attributes* that describes these *objects*. Table 1 shows a formal context in which objects are the applications executed by a user on a tablet, and attributes describe the context in which these applications have been used (details about the dataset used in this article are given in Section 3). For example, *Youtube* has been used at *home*, but also on a *3G* network and in the *morning*. Note that the only possible values in the formal context are 0 and 1, which does not reflect the relative usage of the various applications (indicated in parentheses): *SMS* are associated to *university* and *restaurant* in a similar way in the formal context, although the number of *SMS* sent in each location is very different.

Table 1. Example Formal Context (the values in parentheses correspond to the number of times the application was used in each context).

Applications	University	Restaurant	Parents' home	Home	Transportation	3G	Morning	Coffee break	Lunch break	Afternoon	Evening
Gmail	1 (2)	0	0	0	1	1 (2)	1	0	0	1	0
SMS	1 (80)	1	1 (5)	1 (30)	50	1 (170)	1 (40)	1 (20)	1 (40)	1 (60)	1 (10)
Telephone	0	0	0	1 (3)	1 (2)	1 (5)	1	0	0	1 (2)	1
VDM	0	0	0	1 (4)	1 (2)	1 (5)	1 (4)	0	0	0	1 (2)
Flappy Bird	1	0	0	1	1	1	1	1	0	1	1
Youtube	0	0	0	1	0	1	1	0	0	0	0

Figure 1 represents the Galois lattice generated from the formal context of Table 1. Each concept (cluster) of the lattice groups objects that have attributes in common. For example, this lattice shows that *Flappy Bird*, *SMS* and *telephone* applications have been used in the *afternoon*, in the *evening*, during *transport* and at *home*. These context elements are common to these applications. Note that the concepts become more and more specific when going from the top to the bottom of the lattice, from concepts with many applications sharing few context elements to concepts with few applications sharing many context elements. The lattice provides a clustering of applications according to the context elements they have been associated to. Conversely, context elements are described with regard to the applications associated to them.

One limit of the current approach is that the relative frequency of applications in each context is not reflected in the formal context of Table 1, as the only possible values in this table are 0 and 1. The goal of our work is to take this relative frequency into account in the formal context.

Besides, data collected in a real environment is often complex and heterogeneous. Context information can be represented in very different ways, using symbolic values, real number, intervals, etc. In addition, this information can be inaccurate and lack of precision due to its acquisition method. Many studies propose methods to replace the numeric values (Table 1) by binary values in order to represent the relationships between the objects and the formal context attributes in concept lattice. For instance, [11] use a cut-off threshold to replace values greater than or equal to the threshold values by 1 and the rest by 0. The method in [12] is to set a threshold δ, and then turn the context into a binary one by replacing values less than δ with 0 and 1 the others. However, these thresholds are calculated based on numerical values in the formal context regardless of the semantic link between the attributes. [13] propose a method by intervals; two bounds are defined for each attribute: an upper bound and a lower bound. The formal context is then transformed into binary context by replacing values that are not in the interval with 0 and those that are to 1. Similarly, [14] also proposes an intervals method, in which each numeric attribute is converted into n binary attributes, each one representing a range of values of a numerical attribute. Also, [15] defines a method able to reflect all intervals of possible attribute values. But even with roughly large data, e.g. 200 objects and 200 attributes and a hundred different attribute values, building the concept lattice from the resulting formal context is shown by [16] to be very difficult.

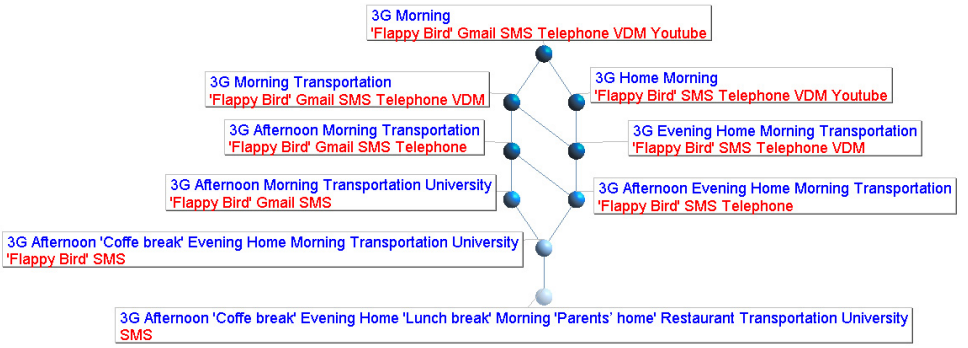

Fig. 1. Galois lattice associated to the formal context of Table 1

3. Proposition and methodology

Our proposal consists in refining formal contexts according to a relevance measure based on the frequency of applications used in the various context elements. We propose two frequency measures that rely respectively on the collected data and on the additional semantic information from context ontology. The obtained frequency values are used to build a refined formal context, following two strategies called high and low, focusing respectively on the most and least frequent context elements and applications. The experiment performed in Section 4 evaluates both frequency measures and both strategies for formal context generation. In the following, we provide more details about each step of our methodology.

3.1. Frequency measures

The first frequency measure *F* is solely based on the number of occurrences of each application in each context element, and does not rely on any external –semantic-information. For example (Table 2), for the *SMS* application of Table 1, each value from the initial data is divided by the total number of occurrences (i.e., 506).

Table 2. Example using frequency measure *F*.

Applications	University	Restaurant	Parents' home	Home	Transportation	3G	Morning	Coffee break	Lunch break	Afternoon	Evening
SMS	80/506	1/506	5/506	30/506	50/506	170/506	40/506	20/506	40/506	60/506	10/506

The semantic frequency measure *Fs* relies on a context ontology, which reflects the various categories of context elements, e.g.: time, location and connection network. The ontology we have used for our experiment is presented in Figure 2. *Fs* divides the number of occurrences of each application in each context by the total number of occurrences of this application in the context elements from the same category in the ontology. For example (Table 3), for the *SMS* application of Table 1, the values related to the location context are therefore divided by 166, which correspond to 80 (*university*) + 1 (*restaurant*) + 5 (*parent's home*) + 30 (*home*) + 50 (*transportation*):

Table 3. Example using semantic frequency measure *Fs*.

Applications	University	Restaurant	Parents' home	Home	Transportation	3G	Morning	Coffee break	Lunch break	Afternoon	Evening
SMS	80/166	1/166	5/166	30/166	50/166	170/170	40/170	20/170	40/170	60/170	10/170

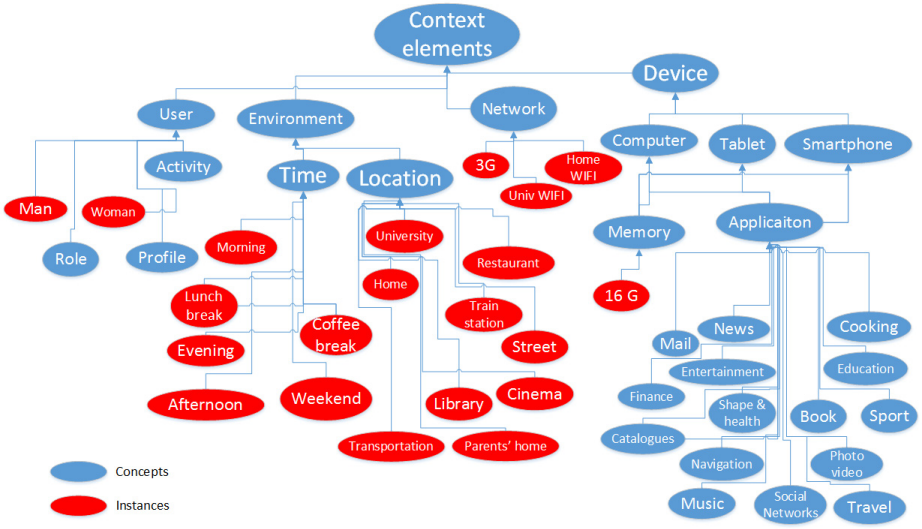

Fig. 2. Context ontology

3.2. Strategies for Formal Context Generation

The average frequency value is computed for each application, i.e., for each line of the table. The high strategy consists in assigning the binary value of 1 to the most frequent context elements of each application. In other terms, for each application, each value greater than a given *high_threshold* for this application is associated to 1, and the others to 0. We define the *high_threshold* as follows:

high_threshold = *average_frequency* + β * *frequency_standard_deviation* (β is a positive or null float)

The higher β is, the more restrictive the refinement is, i.e., the fewer values are assigned to a "1" in the resulting formal context, and therefore taken into account for the Galois lattice generation. For example (Table 4), if we consider the *SMS* application and a value of 0 for the β parameter (bearing in mind that we have tested other values of β), each value greater than the *high_threshold* (0.27) is assigned to 1 and the others to 0. The corresponding line in the new formal context is:

Table 4. Example with a *high_threshold* and $\beta=0$.

Applications	University	Restaurant	Parents' home	Home	Transportation	3G	Morning	Coffee break	Lunch break	Afternoon	Evening
SMS	1	0	0	0	1	1	0	0	0	1	0

On the other hand, the second strategy, called low strategy, assigns 1 to the frequencies that are lower than a *low_threshold* for each application, where:

*low_threshold = average_frequency - β * frequency_standard_deviation* (β is a positive or null float)

For the *SMS* application, with β=0, the resulting (Table 5) formal context with the low strategy is:

Table 5. Example with a *low_threshold* and $\beta=0$.

Applications	University	Restaurant	Parents' home	Home	Transportation	3G	Morning	Coffee break	Lunch break	Afternoon	Evening
SMS	0	1	1	1	0	0	1	1	1	0	1

The following Section describes our experiment using these strategies and measures.

4. Experiment

We have evaluated our proposal on data collected from a questionnaire filled by 28 master students of our university. The goal of this survey was to know in which context and on which device they had been using applications (such as social networks, games, emails, etc.) during a week. The students have indicated how many times they have used each application in various contexts (the context elements were not pre-defined: they have been filled in by the students themselves) using a questionnaire instead of an application for automatically recognize context elements allows us to consider context elements that are not predefined by the observation mechanism, offering more flexibility for the volunteers students. The ontology presented in Figure 2 has been built in order to reflect all the context elements that have been mentioned in the questionnaires.

4.1. Evaluated strategies

In this experiment, we have evaluated 4 strategies for formal context generation:

- Strategies 1 and 2 are high strategies that focus on the most frequent context elements for each application. Both strategies differ by the frequency measure: strategy 1 does not use any semantic information, whereas strategy 2 relies on the context ontology (*Fs* measure).
- Strategies 3 and 4 are low strategies that focus on least frequent context elements for each application. Strategy 3 uses frequency *F*, whereas strategy 4 uses *Fs*.

Table 5. Strategies description.

	Most frequent context	Least frequent context
First frequency (*F*)	S1	S3
Semantic frequency(*Fs*)	S2	S4

Each strategy has been tested with various values, more or less restrictive, of the refinement threshold, i.e., associated to different values of β parameter (respectively 0, 0.25, 0.5, 0.75, 1 and 1.5). We should mention that the

values of β were not chosen randomly, we have begun with the value of zero for β and then we increased the value of β slightly. Further on we analyzed the results after using the different values, stopping when the number of relations was very limited to extract and interpret the results.

4.2. Comparison of frequency measures

We first compare the various strategies for the least restrictive refinement threshold, i.e. *high_threshold=low_threshold=average_frequency* (for both frequency measures). Strategies using the same frequency measure can be compared, i.e., strategies 1 and 3 on the one hand, and strategies 2 and 4 on the other hand. As we could expect, the frequencies obtained with the measure that is based on the context ontology are more relevant than the ones obtained from the one that relies on no semantic information. For example, the *3G* context element does not appear as very significant among all other context elements in strategy 1, whereas the ontology indicates that it is indeed the only context element related to the type of network connection in this dataset, and therefore it should not be neglected. However, our experiment shows that even without using an ontology-based frequency measure, our methodology for building formal contexts that reflect the impact of context elements on applications still provides interesting results with regard to resulting Galois lattices.

In the following, for space reasons, we focus only on ontology-based strategies (2 and 4), that have provided optimal results.

4.3. Comparison of refinement strategies

With β=0, both high and low strategies lead to larger Galois lattices; it means that the refined formal context generates more concepts, as illustrated in Figure 3 for the high strategy. This lattice is more interesting than the original lattice of figure 1 for recommendation and personalization purposes, as more clusters of applications and context elements are generated, leading to a finer-grained classification. In Figure 1 for example, *Flappy bird, SMS, telephone, VDM* and *Youtube* applications have in common the context elements *3G, home* and *morning*, whereas they are separated into 2 concepts in Figure 3: *Gmail, VDM* and *Youtube* are associated to *3G* and *morning*, and *Flappy bird, telephone, VDM* and *Youtube* are associated to *3G* and *home*. It means that after the refinement strategy, *Flappy bird* and *telephone* are no longer associated to the morning (and will therefore not be considered as very likely to be used in the morning).

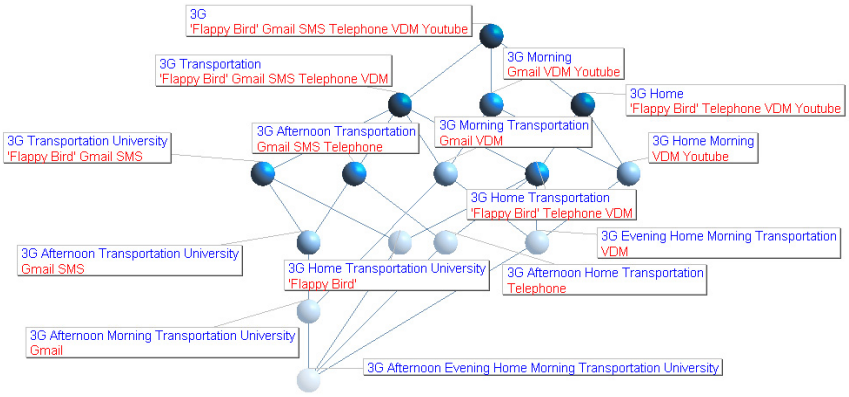

Fig. 3. Galois lattice obtained with strategy 3 (ontology-based frequency measure and *high_threshold*), with β=0

We also notice that some context elements have disappeared from the lattice, as they have not sufficient impact on the user applications: *coffee break*, *restaurant*, *parents' home* and *lunch break*. It is interesting to remark that the number of concepts in the lattice increases despite the disappearance of 4 context elements. This shows that the precision of the obtained results has indeed increased significantly with regard to the initial lattice.

The formal contexts obtained with higher values of the β parameter, i.e., with stronger refinement conditions, generate smaller Galois lattices than the ones obtained with β=0 (for both high and low strategies). For our

recommendation and personalization purposes, we therefore choose the lower value of β as large lattices provide a finer granularity.

We will not show the lattice obtained for the low-strategy β=0 and for space reasons, but it stresses the context elements which are rarely associated to the applications of a given user. For example, one of the obtained concepts associates *telephone* and *Youtube* to *coffee break, evening, lunch break, parents' home, restaurant* and *university*. It means that this student has (barely) not used these applications in these contexts. This information is very valuable to understand application usages. Service and application designers may want to exploit it to try and find incentives to motivate users to use them in wider contexts.

5. Conclusion and perspectives

In this paper we have proposed strategies and relevance measures for data refinement in Pervasive Information Systems to assess the relevance of contextual data through Formal Concept Analysis. We have proposed a methodology to apply refinement strategies with different frequency measurements on contextual elements. The lattice construction based on refined formal contexts has allowed us to identify sets of relevant contextual elements. We have applied our proposal to a case study based on real data. We should mention that the collected data was used to run an offline analysis to evaluate and validate the potential results. However, an online analysis will be done in our future research, considering that the interest of such analysis during the runtime. In future work, we will use more restrictive threshold values, knowing that the best results were obtained when for β=0. Another perspective is to analyze and compare the results of all students, by doing an automatic interpretation taking in account all different strategies and values. Currently, we want to a find relevant new interpretations on each Galois lattice based on application usage according contextual elements, these interpretations are based on statistics and calculations through lattice. Also we want to compare the entire Galois lattice representing the results of our refinement strategies. We use the dataset of users to build our strategies. These data represent the different applications with different user device and the contextual elements.

References

1. Kourouthanassis, P.E., Giaglis, G.M., Vrehopoulos, A., Enhancing the user experience with pervasive information systems. *International Journal of Information Management*, 2008; **27**: 319-335.
2. Dey, A.K., Understanding and using context. *Personal and ubiquitous computing*, 2001; **5**(1):4-7.
3. Jaffal, A., Kirsch-Pinheiro, M., Le Grand, B., Unified and Conceptual Context Analysis in Ubiquitous Environments. *8th International Conference on Mobile Ubiquitous Computing, Systems, Services and Technologies*, 2014, p. 48-55.
4. Baldauf, M., Dustdar, S., Rosenberg, F., A survey on context-aware systems, *Int. J. of Ad Hoc and Ubiquitous Computing*, **1 2**(4): 263-277.
5. Greenberg, S., Context as a dynamic construct. *Human-Computing Interaction*, 2001; **16**(2-4):257-268.
6. Fayyad, U., Piatetsky-Shapiro, G., Smyth, P., From data mining to knowledge discovery in databases. *AI Magazine*, 1996; **17**(3): 37-54.
7. Chen, M.S., Han, J., Yu, P.S., Data mining: an overview from database perspective. *IEEE Transactions on Knowledge and data Engineering*, **8**(6): 866-883.
8. Ramakrishnan, A., Preuveneers, D., Berbers Y., Enabling self-learning in dynamic and open IoT environments. *5th Int. Conference on Ambient Systems, Networks and Technologies (ANT-2014)*, Procedia Computer Science, Elsevier, 32: 207-214.
9. Priss U., Formal Concept Analysis in Information Science. *In: Blaise, C. (ed.) Annual Review of Information Science and Technology*, 2006; **40**, p. 521-543.
10. Wille, R., Formal Concept Analysis as Mathematical Theory of Concepts and Concept Hierarchies. *In: B.Ganter et al., Formal Concept Analysis, Springer-Verlag*, 2005, p. 1-33.
11. Pensa, R.G., Leschi, C., Besson, J., Boulicaut, J.F., Assessment of discretization techniques for relevant pattern discovery from gene expression data. *4th ACM SIGKDD Workshop on Data Mining in Bioinformatics,* 2004, p. 24-30.
12. Ma, J., Zhang, W.X., Cai, S., Variable threshold concept lattice and dependence space. *Proceedings of the Third International Conference on Fuzzy Systems and Knowledge Discovery*, 2006, p.109-118.
13. Zhou, W., Liu, Z., Zhao, Y., and Xie, Z., Clustering-based reduction algorithm on the structure of fuzzy concept lattices. *5th International Conference Formal Concept Analysis, ICFCA07*, 2007, p. 131-145.
14. Kaytoue, M., Duplessis, S., Napoli, A., Using formal concept analysis for the extraction of groups of co-expressed genes. *In : An, L.T.H., Bouvry, P., Tao, P.D., Eds., MCO,CCIS*, *Springer*, 2008, p. 439-449.
15. Ganter, B.,Wille, R., Formal Concept Analysis. *Springer, mathematicalfoundations edition*, 1999.
16. Kaytoue, M., Duplessis, S., Kuznetsov, S.O., Napoli, A., Two FCA-Based methods for mining gene expression data. *In: Ferré, S., Rudolph, S., (Eds.), Formal Concept Analysis, Springer*, 2009, p. 251-266.