Urban Knowledge Extraction, Representation and Reasoning as a Bridge from Data City towards Smart City

Jaime de Miguel-Rodríguez
Depto. Ciencias de la Computación e Inteligencia Artificial
Universidad de Sevilla
Sevilla, Spain
Email: demiguel.jaime@gmail.com

Juan Galán-Páez
Depto. Ciencias de la Computación e Inteligencia Artificial
Universidad de Sevilla
and
Datrik Intelligence S.L.
Sevilla, Spain
Email:juangalan@us.es

Gonzalo A. Aranda-Corral
Depto. Tecnologías de la Información
Universidad de Huelva
Palos de La Frontera, Spain
Email:garanda@us.es

Joaquín Borrego-Díaz
Depto. Ciencias de la Computación e Inteligencia Artificial
Universidad de Sevilla
Sevilla, Spain
Email:jborrego@us.es

***Abstract*—Urban Data management represents a major challenge in the field of Smart Cities. Its understanding is essential for the development of better smart services, which are a persistent demand in urban policies. From all the sources of data available, those that involve a collective processing of urban information (by the citizens or other collectives) deliver in fact, useful insights into social perception. Such is the case, for example, of data collected from mobile networks. Prior to the design of socio-technical artifacts in cities, it seems important to extract the qualitative and quantitative opinions, sentiment and feedbacks present in these data. In this paper we present three solutions for mining these contents through Knowledge Extraction methods, as a previous step to the prospection of new smart services.**

I. Introduction

We are living an increasing availability of data and information layers that can be processed jointly. Information coming from individuals, communities, businesses and institutions, and enclosing knowledge about the complex systems in which these sources interact. Data reflects not only relationships between users but also our behavior as a community from different facets: both current and historical, the technological point of view and also the cultural, economic, social or techno-social view. Also, this information satisfies the three V's characterizing the projects that can be included within the paradigm of Big Data: Volume, Variety and Velocity.

Until relatively recently the majority of such information was approached from a one-dimensional perspective, that is, its analysis was usually limited to one, or sometimes two, dimensions in terms of the relationships between the "micro" components of the systems under study. The main reason behind this fact is that the aim of these studies was defined and fixed beforehand with more or less precision. Because of the large variety of possible approaches to the analysis of these systems, their information is traditionally treated, analyzed and interpreted from a specific and strategic point of view, tailored for the results that can be obtained: forecasting, learning, classification, etc.

This restriction on data analysis makes it harder to extract wisdom from the information available, thus, limiting the potential of Smart-City application tools that make use of it. However, the extraction of knowledge from a corpus of specific concepts of a Big Data system allows to characterize the essential properties of both the system and its dynamics. In addition, knowledge extraction provides with the necessary means to reason about the system at a macro level. This knowledge is the basis to empowering multi-purpose tools by streamlining the specialization of the systems to provide specific services.

There exist two viewpoints for urban digital information. On the one hand, there exists the vision of *Data City*, a place where data is collected, being the basis of the Geospatial Cyber Infrastructure (GCI) among other information ecosystems [17]. And on the other hand, the city can be considered as an interactive system, where social (real and virtual) networks interplay, including local interactions (among users, between users and concrete places [11]) and global ones (between citizens of different cities).

Although the natural evolution of the GCI points towards a constellation of smart urban services, at the present moment only a few cities offer such opportunity, ie. a city implementing a GCI advanced and flexible enough to provide data to companies, citizens and enterpreneurs for deploying innovative applications. It is usual to broaden the scope of data sources by

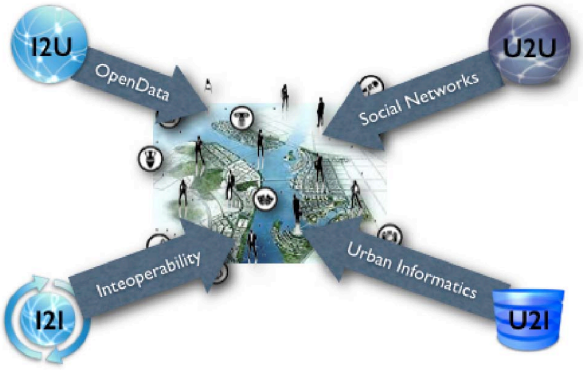

Fig. 1. Information Flows in Digital Cities

looking into to classical web and data services (with variable accessibility) and data collected from the WWW (of variable quality).

Clearly, these sources have limitations that do not allow a perfect digital image of the city, but on the other hand, they do create opportunities to analyze important aspects of urban behavior. This scenario has been greatly enhanced with the efforts put on *open data* in many cities, which allow to incubate R & D initiatives.

There exist four types of information flows which can be considered in order to analyse how digital information can be used in simulation: U2U (users to users), I2U (institutions to users), U2I (users to institutions) and I2I (institutions to institutions) (see Fig. 1).

Each kind of flow has specific features that affects its explotation (transformation, extraction, etc. See Fig. 2). In general, the falling cost of data coverage and a greater availability of map software have led to demonstrably rapid advances [20].

The aim of the paper is to show how the extraction of knowledge from these flows greatly benefits the design of smart services which are strongly related with the social dimension of cities.

II. Semantic Web technologies and the Social dimension

The Semantic Web (SW) could be useful to organize and reason with urban knowledge by using ontologies [10]. In some cases, the available urban information provides guidelines to their creation [8] whilst in others, the concepts hidden within the data are not clear. In this case, Knowledge Extraction (KE) methods are necessary. In fact SW technologies may represent tight restrictions to analyze emergent semantics from U2U flows. In fact, U2U flows can represent truly new institutions on the own social movement or networks which produce the information (see e.g. [6]).

From the *Data City* point of view, the situation is different: accessing and processing are more transparent and trustworthy. In some cases, the available urban information provides a number of guidelines for the creation of semantic infrastructures [8]. In this case *Geodemographic information systems* represent a kind of business tool for interpreting data, consisting mainly of a demographic database, digitized maps, and software. They are widely used for several applications [9], by providing useful semantic insights for the construction.

In the context of the Smart-City information ecosystem, the realm of social knowledge brought forward by a geodemographic ontology would influence all information collection, interpretation and feedback processes within the Urban Informatics scope, as for example in the city management, by leading the specialization of decisions and applications.

In Fig. 3 the transformation from information-based life cycle in cities into the knowledge-based one is depicted (the top cycle is from http://www.cityofsound.com/blog/2008/08/two-or-three-re.html). The semantic layer allows by means of their use on SW technologies (as ontologies) to instantiate information about the own urban information design in Knowledge. In this way, decisions about the city are argumentative, tested and their trust can be evaluated. On the other hand, semantic processing on semantic data (geodemographics, about human mobility, traffic, etc.) can be supported by sensor information, and reciprocally, semantic processing can induce to reorganize sensors distribution and interaction.

With this approach, the life cycle of knowledge in smart cities (including acquisition, verification, documentation and decision-making) can be enriched with semantic processing of data, besides sociodemographic ontologies [8] or similar formal artifacts. The value added by semantic technology allows us to mediate via (high level) reasoning with the processed knowledge. Of course, this strategy does not exclude the fact that data coming from collaborative practices or *crowdsourcing*, is an important feature of the *Interactive City*, which is often under-used (for example in urban policies and planning).

In the *Interactive City* point of view, concepts hidden within the data are not clear. Data is not collected and formatted for specific uses, for example when we need to study the relationship among digital information and citizens [15]. Here citizens produce and consume data with relative freedom and they do not care how social perception about the specific topic could be represented in a standard and explicit way. That is to say, semantic cities do not emerge from P2P urban digital information in a direct way. For example, it is necessary to carry out *concept mining* tasks in order to understand how citizens manage (process) the information. In this case, Knowledge Extraction (KE) methods are necessary to bridge the semantic gap between data and socioeconomic perception.

III. Extracting emergent knowledge by using Formal Concept Analysis

The method for Knowledge Extraction is based on Formal Concept Analysis (FCA) [14]. According R. Wille, FCA mathematizes the philosophical understanding of a concept as a unit of thoughts composed of two parts: the extent and the intent [14]. The extent covers all objects belonging to this concept, while the intent comprises all common attributes valid for all the objects under consideration. It also allows the

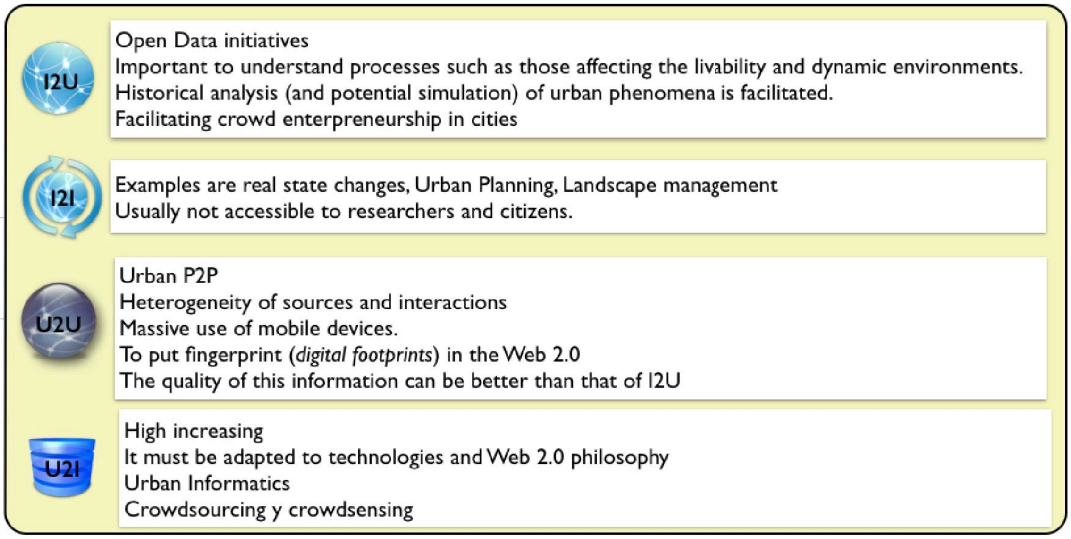

Fig. 2. Main features of information flows in Digital Cities

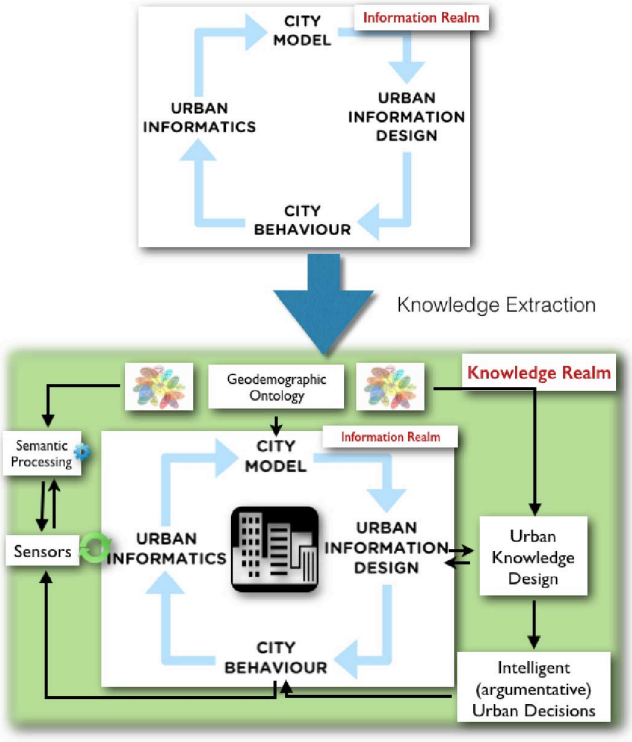

Fig. 3. From Information Realm to Knowledge in the urban information ecosystem [8].

computation ofconcept hierarchies (*concept lattices*) from data tables. In the case of great amount of information associated to complex systems, the concept lattices are complex semantic networks with specific topological properties [2].

FCA also provides reasoning tools which can be used to re-organize, enrich [4] and even predict the evolution of the system [3]. It also provides powerful semantic tools for classification, data mining and KE and Discovery (KD). Among these tools, particularly interesting are concepts extraction and organization, and implication basis. The last one represents a sound approach to rule extraction for classification. This task is a significant issue in KD where FCA applications in the field of Soft Computing have been implemented (see for example [18]).

Formal Concept Analysis (FCA) can be used to organise knowledge and extract new concepts from rear data:

- Digital Information describing features of cities hide (qualitative) human behaviours and beliefs.
- The extraction of knowledge from digital footprints helps to understand how citizens live and work within the city.
- There exist hidden ontologies useful in the (bounded) reasoning on cities and their structure.

The methodology applied (see Fig. 4) is based on the processing of data to obtain a formal context, which represents the global knowledge (called *Monster Context*). Afterwards, the concept lattice is computed, consisting of a complex semantic network that presents a complex topology [2]. The second stage involves the selection of a number of features and thresholds (in order to discretize data) which are used for producing Knowledge bases in reasoning tasks. Concept lattices represent weak ontological structures which can be specified for different urban spaces, cities, etc.

This methodology was applied in three cases with different levels of generality:

- In [13] is applied to the Housing Market, showing (emergent) semantic patterns on the social perception of housing values in different urban areas of the city. By means of the use of association rules (associated to semantic networks produced by FCA) it is possible to understand

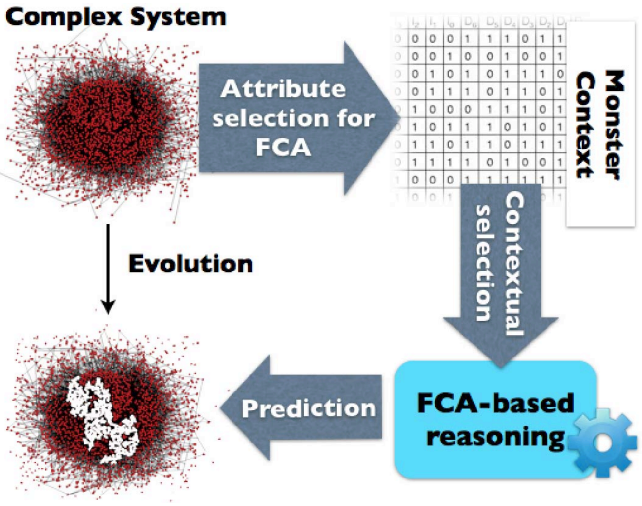

Fig. 4. FCA-based model for qualitative reasoning with Complex Systems from observations

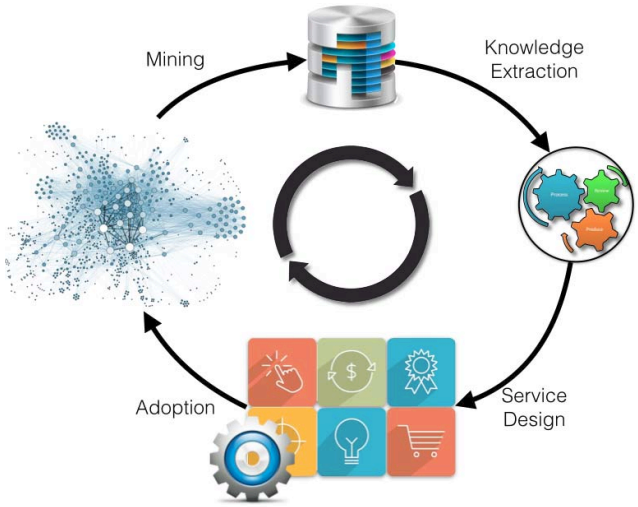

Fig. 5. Design from Urban Knowledge

the qualitative nature of the distinct perceptions, a nature based on emergent semantics instead of purely statistical features.
Although the flow has technological nature as a I2U one, the information is crowdsourced among users of housing market social networks; that is, it is U2U but mediated and constrained by Social Web platforms.

- In [12] is used to design a Multiagent model of a *micro* urban nature phenomena: (qualitative based) behavior mining from location-based data about pedestrian activity. Information is mined from user activities (for example, from sensors or their own mobile devices).
- In [1] the method is applied to estimate the quality of information provided by institutions about the city by comparing emergent semantics of different counties (open government data). This case is concerned with socioeconomic information curated by institutions (I2U).

IV. Product/Service Design from FCA-based Methodology

The aforementioned cases foster the design of smart city applications and services. The life cycle (see Fig. 5) is based on four stages:

- *Information Extraction* from physical (sensors) and digital (e.g. WWW) sources.
- *Knowledge Extraction* and processing (by using FCA-based techniques, in our case). Data integration, data cleansing and feature selection are involved, among other activities.
- *Service Design and deployment*, based on the exploitation of Knowledge by means of intelligent methods and interfaces. The activities depend on which channel information will be used and which will be the users.
- *Adoption Phase*. This stage is beyond the scope of this paper. In fact it can be considered as an activity within the Information Service Evaluation Activity: How do users adopt, use and accept those services? Do the information services exercise influence over the users' behavior? How do such services diffuse into society? [16]

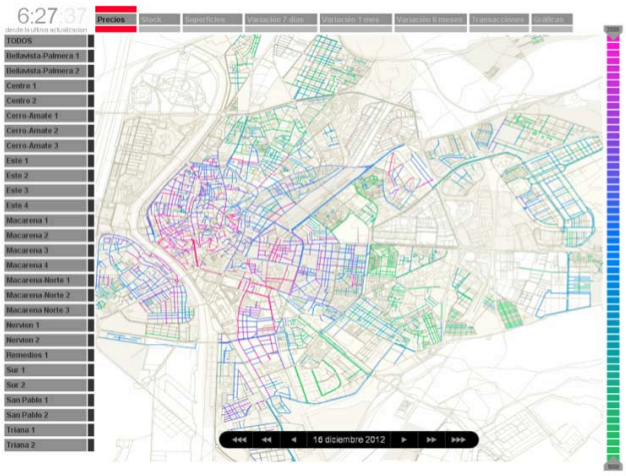

Fig. 6. Self-City

A. Case 1: Estimating social perception on housing value in cities

- Information Extraction: This case needs of specific wrapper-based information extraction systems, as for example *Self-City* (see Fig. 6). Self City enables the incorporation of real time urban information (in housing markets) in a meaningful way, striving to make dynamic data both efficient and fundamental to city management and even decision-making.
Currently, Self-City manages only digital information. Such information can be visualized, but a sound compre-

hension of its dynamics and conceptual structure requires a semantic processing. The understanding of the full data set needs a semantic interpretation of the concepts involved in this complex system.

- Knowledge Extraction: Semantics provided by FCA allow to estimate both social opinion about housing prices and their location. In Fig. 8 a social semantic pattern on house price in the neighboring area of a main avenue in the city of Seville is depicted.
  Contexts are built from the Self-City information flow, as for instance temporal and spatial information on for-sale housing: objects are for-sale homes and attributes represent a qualitative description of these homes at a given moment; description of the item (i.e. price, dimensions, environment, etc.) and description of the item evolution (i.e. price changes, environment evolution, etc.). Thresholds for attributes have to be selected.
  From this context, the concept lattice is computed. In Fig. 7 the source is a collection of 6000 (approx.) for-sale homes in the city of Seville. Also subareas of Seville can be considered (i.e. streets, zones, etc.) for a more detailed analysis. A more precise analysis of conceptual differences can be provided by means of logical reasoning with association rules related to the lattice [5]. Also, it is possible to consider how qualitative patterns are reproduced within the city. The study of this kind of semantic structures allows, for example:
  - *Discovering new concepts relevant for housing markets*. The discovering comes from the analysis of the Concept Lattice. The concept lattice for a particular city area allows both understanding and comparing different city areas and neighborhoods.
  - *Comparing socio-economics contexts*. The estimation of similarity among concept lattices from different cities allows comparing socioeconomic contexts. Fig. 8 shows the similarity between the main streets of a district and a specific Avenue. Patterns have to be interpreted and studied by experts in urban planning.
- Service design proposal: Mainly two-fold. On the one hand Self City can be enhanced with semantic information by means of intelligent interfaces. On the other hand it allows to design a Location-Based Service (as an app, for example) which provides with the social perception of housing assets in the neighborhood where the user is in this moment or where he/she lives in.
  Lastly, semantic enhancing of Self City could provide information to urban authorities about sustainability issues, as for example seminal gentrification processes in concrete areas or contiguity effects (cf. [19]).

B. Case 2: Exploiting pedestrian behavior in streets for smart mobility

The aim of this case is to show how to exploit knowledge extracted from observations (of real or artificial systems) to study and explain -in a qualitative formalism- pedestrian behavior [12].

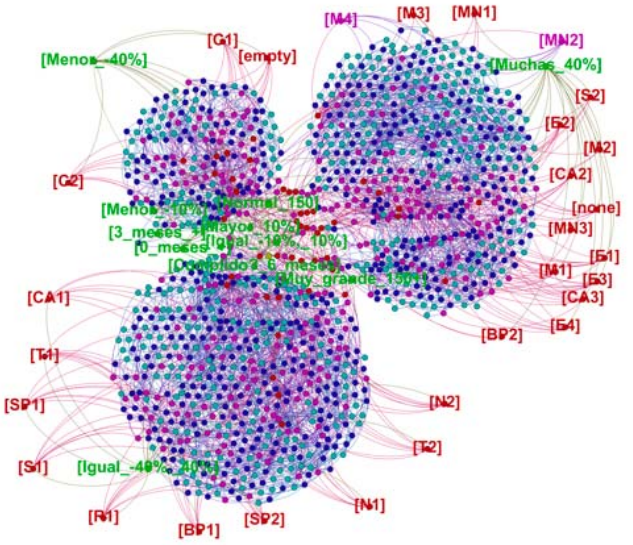

Fig. 7. Concept Lattice on housing from WWW data about Seville

- Information Extraction: Information from sensors and GPS-locations of users (from their smartphone). The current datasets are not recollected, they are synthetic. However it is not difficult to adapt the prototype to real data (discretization process).
- Knowledge Extraction: The result of this process is, itself, a knowledge based system that is also useful to simulate the source system.
  This work is based on the intensive use of FCA which provides mathematical tools for detecting qualitative concepts, useful in the phenomenological reconstruction of Complex Systems [2]. In this case, associated to pedestrian mobility (Fig. 9).
- Service Design: By using this system as a deliberative module for agents, we have implemented a general simulation framework for natural and artificial models of mobility.
  Also, simulated models provide useful information to citizens with reduced mobility for the task of deciding alternative routes or behaviors.

C. Case 3: Exploting Real Time Government Information

Semandal is a platform that tries to apply all concepts about Semantic Web and Open Government Data on the network of municipalities of Andalucía, Spain. It is an example of second-level social network built on the socioeconomic complex system.

- Information Extraction: Semandal extracts the information from traditional web pages.
- Knowledge Extraction: Semandal transforms knowledge by means FCA. To transform information into knowledge

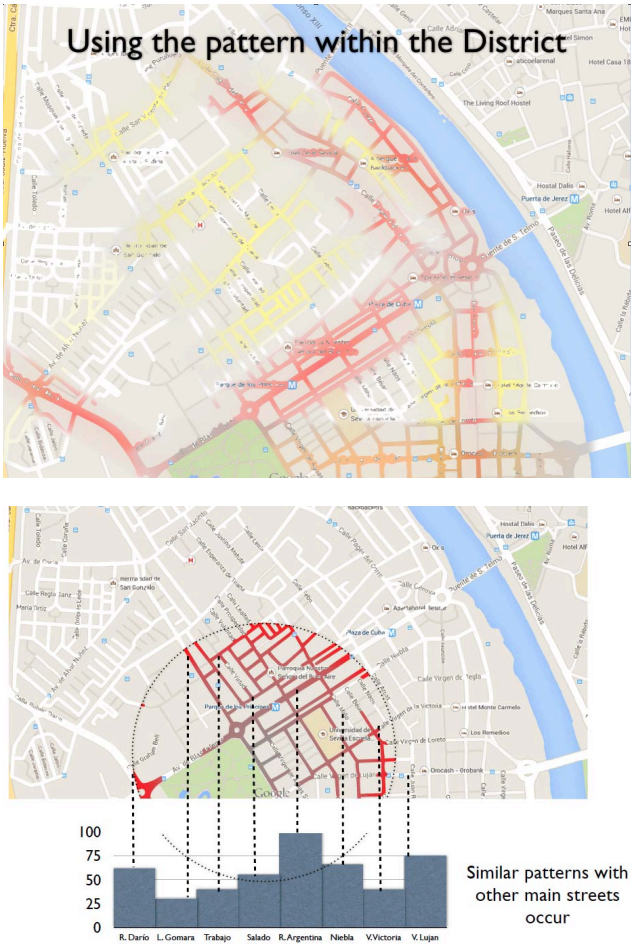

Fig. 8. Semantic (social) Patterns of Housing Market in a district from Seville (top) and similarity measurements (bottom)

we use FCA as a non supervised clustering technique which can construct concept lattices and sets of rules which represent information. Semandal's architecture is depicted in Fig. 10.

- Service design: Knowledge is published by means of an API and a structured format (machine-readable). Even this, Semandal provides a mobile app to let users access this information (see Fig. 11).

V. Conclusions

In this work a number of urban Knowledge-Based applications and services are presented. The idea is to show how Knowledge Engineering techniques can enhance city services towards smart services. Each example uses information sources of different nature, and the product/service designed attempts to exploit emergent semantics. Each stage of the methodology involves a number of Data Science activities, Knowledge Engineering methods and Software Engineering to obtain added value.

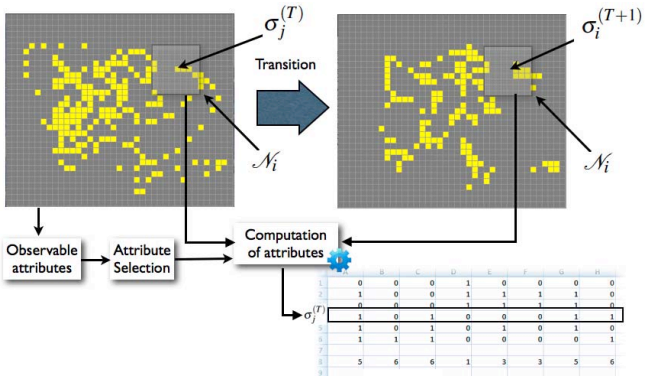

Fig. 9. FCA-based modeling of pedestrians behavior from [12]

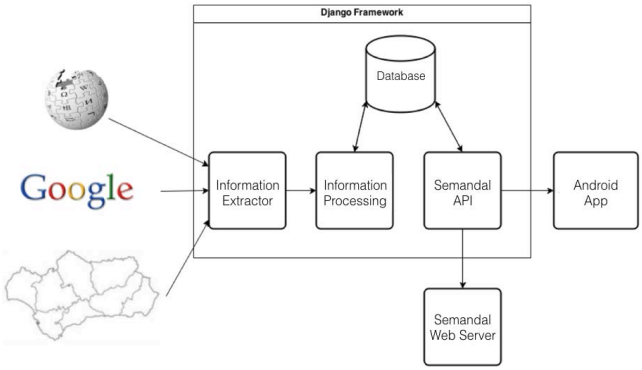

Fig. 10. Semandal's Architecture

Future work is oriented to acquire better urban knowledge mined from citizens sentiments and opinions. As it was pointed out in [7], this kind of analysis could be an interesting source of citizen's feedback (see also [15]). For example, it can be used to forecast real estate evolution.

VI. Acknowledgements

Partially supported by TIN2013- 41086-P (Spanish Ministry of Economy and Competitiveness), co-financed with FEDER funds.

References

[1] D. Albendín-Moya, G. A. Aranda Corral, J. Borrego Díaz. A. Cantó-Vicente. Semandal: Extracting knowledge and data from city councils, in Proc. 7th European Symposium on Computational Intelligence and Mathematics, pp. 180-85 (2015).

[2] G.A. Aranda-Corral, J. Borrego-Díaz, J. Galán-Páez, On the Phenomenological Reconstruction of Complex Systems—The Scale-Free Conceptualization Hypothesis, Systems Research and Behavioral Science 30(6): 716-734 (2013).

[3] G. A. Aranda-Corral, J. Borrego-Díaz, J. Galán-Páez, Complex Concept Lattices for Simulating Human Prediction in Sport. J. Systems Science and Complexity 26(1):117-136 (2012)

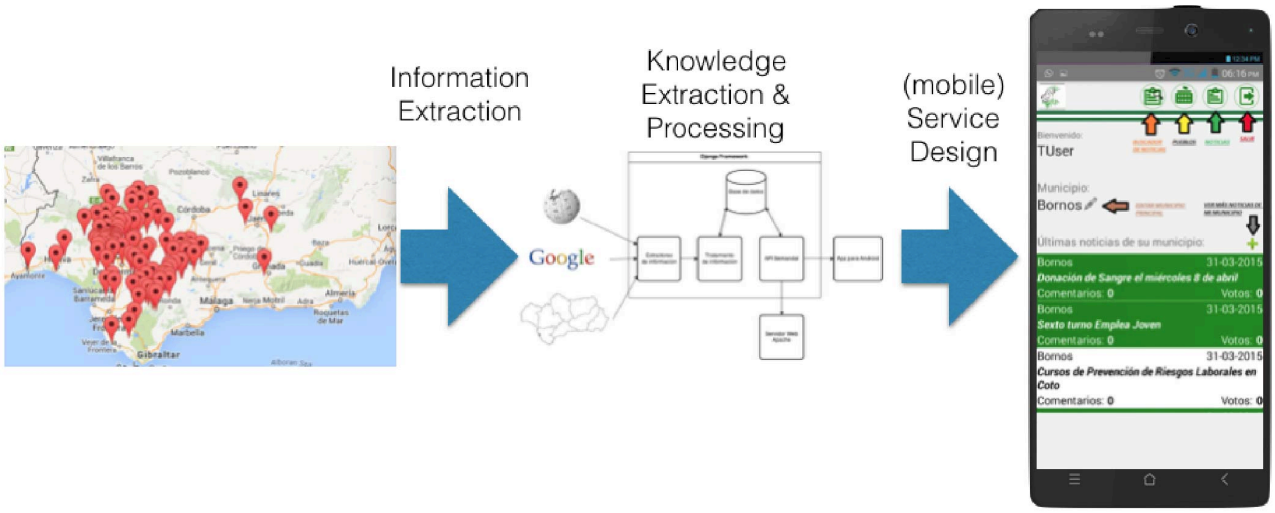

Fig. 11. Semandal's Knowledge-based process

[4] G. A. Aranda-Corral, J. Borrego-Díaz, J. Giráldez-Cru: Agent-mediated shared conceptualizations in tagging services, Multimed Tools Appl. 65(1):5-28 (2013).

[5] Aranda-Corral, G. A., Borrego-Díaz, J., Galán-Páez, J.: Confidence-based reasoning with local temporal formal contexts, Proc. 11th Int. Conf. Artif. Neural Networks, LNCS vol. 6692, pp. 461-468. Springer (2011).

[6] D Beraldo, J Galan–Paez. The #OCCUPY network on Twitter and the challenges to social movements theory and research. Int. J. Electr. Gov. 6 (4), 319-341 (2014)

[7] J. Borrego-Díaz, J. Galán-Páez, Discovering new sentiments from the social web. Proc. Collective Intelligence 2014 arXiv preprint arXiv:1407.0374, abs/1407.0374 (2014).

[8] J. Borrego-Díaz,A. M. Chávez-González, M.A. Martín-Pérez, J.A. Zamora-Aguilera, Semantic Geodemography and Urban Interoperability, Metadata and Semantics Research, Comm. Computer and Inf. Science Volume 343, 2012, pp 1-12

[9] R. Burrows, N. Ellison, B. Woods, Neighbourhoods on the net: The nature and impact of internet-based neighbourhood information systems. Joseph Rowntree Foundation, 2005.

[10] G. Falquet, C. Métral, J. Teller, C. Tweed (eds.) , Ontologies in Urban Development Projects Advanced Information and Knowledge Processing Volume 1, 2011

[11] Foth, M. (Ed.) Handbook of Research on Urban Informatics: The Practice and Promise of the Real-Time City. Information Science Reference, IGI Global (2009).

[12] J. Galán Páez, G. A. Aranda-Corral, J. Borrego Díaz, Synthetising Qualitative (Logical) Patterns for Pedestrian Simulation from Data. To appear in Proc. SAI Intelligent Systems Conference 2016 September 21-22, 2016 London, UK.

[13] J. Galán Páez, J. Borrego-Díaz, Jaime de Miguel-Rodríguez: Extracting emergent knowledge about the socioeconomic urban contexts. UbiComp/ISWC Adjunct 2015: 1571-1574

[14] B. Ganter and R. Wille, *Formal Concept Analysis - Mathematical Foundations*, Springer-Verlag, 1999.

[15] Kukka H, Kostakos V, Ojala T, Ylipulli J, Suopajärvi T, Jurmu M, Hosio S. (2011) This is not classified: everyday information seeking and encountering in smart urban spaces Personal and Ubiquitous Computing, 2011

[16] Laura Schumann, Wolfgang G. Stock Webology, The Information Service Evaluation (ISE) Model, 11(1):1-20 (2014).

[17] Yang, C., Raskin, R., Goodchild, M., and Gahegan, M. 2010. "Geospatial cyberinfrastructure: Past, present and future." Computers, Environment and Urban Systems, 34 (4): 264–277.

[18] J. Yu, C. Li, W. Hong, S. Li, D. Mei, A new approach of rules extraction for word sense disambiguation by features of attributes, App. Soft Comput. 27: 411-419 (2015).

[19] Changshan Wu and Rashi Sharma, Housing Submarket Classification: the Role of Spatial Contiguity, Applied Geography, 32 (2), 746-756.

[20] Black & Veatch consulting. 2016 Strategic Directions: U.S. Smart City/Smart Utility Report, 2016 https://pages.bv.com/SDR-SmartCitySmartUtility-DL.html