The Digital Emerging and Converging Bits of Urbanism

Crowddesigning a live knowledge network for sustainable urban living

Flora Dilys Salim[1], Jane Burry[2], David Taniar[3], Vincent Cheong Lee[4], Andrew Burrow[5]
[1,2,5] Spatial Information Architecture Laboratory (SIAL), RMIT University, Melbourne VIC, Australia, [3,4] Faculty of Information Technology, Monash University, Clayton VIC, Australia
[1]http://florasalim.com,[2]http://www.sial.rmit.edu.au/People/jburry.php, [3]http://users.monash.edu.au/~dtaniar/, [4]http://www.infotech.monash.edu.au/about/staff/Vincent-Lee, [5]http://www.sial.rmit.edu.au/People/aburrow.php
[1]flora.salim@rmit.edu.au, [2]jane.burry@rmit.edu.au, [3]david.taniar@monash.edu, 4vincent.cs.lee@monash.edu, [5]andrew.burrow@rmit.edu.au.

Abstract. *Data is ubiquitous in our cities. However, designing a knowledge network about our cities is an arduous task, given that data sensed cannot be used directly, human significance must be added. Adding human significance can be achieved via an automated "expert system (ES)" in which domain expert knowledge are stored in a knowledge-based repository. The domain expert knowledge is matched with the corresponding data to derive specific inference which can aid decision making for urban stakeholders.This requires amalgamation of various interdisciplinary techniques. This paper presents a survey of existing technologies in order to investigate the emerging issues surrounding the design of a live knowledge network for sustainable urban living. The maps and models of the existing infrastructure of our cities that include a wealth of information such as topography, layout, zoning, land use, transportation networks, public facilities, and resource network grids need to be integrated with real-time spatiotemporal information about the city. Public data in forms of archives and data streams as well as online data from the social network and the Web can be analyzed using data mining techniques. The domain experts need to interpret the results of data mining into knowledge that will augment the existing knowledge base and models of our cities. In addition to the analysis of archived and streamed data sources from the built environment, the emerging state-of-the-art Web 2.0 and mobile technologies are presented as the potential techniques to crowddesign a live urban knowledge network. Data modeling, data mining, crowdsourcing, and social intervention techniques are reviewed in this paper with examples from the related work and our own experiments.*

Keywords. *Crowdsourcing; Knowledge Discovery; Mobile and Ubiquitous Computing; Urban Modeling; Spatial Interaction; Social Networking; Web 2.0*

Background

The growth of personal and mobile computing, wearable devices, sensors, and digital artifacts in our built environment has transformed spatial interactions in urban cities. The profusion of streams of data in our urban environment provides short burst of patterns that can lead to useful hidden attributes being discovered for use to improve sustainable developments. Archived and real-time sensor data from the urban context provides the potential for monitoring the current operating conditions of the city. Lessons can be drawn from existing building projects that have good energy performance ratings, demonstrate sustainable building operations, or accommodate satisfied occupants. Solar and energy analysis can be performed on the existing land use of the built environment. Residential and office buildings are now installed with smart meters which stream real-time energy use data. Coupled archived socio-economic data and data from energy simulations of the built environment, live sensor data after matching with expert domain specific knowledge can provide information that is pertinent to discovering patterns of sustainable urban living.

Inherently, the issues of transportation are closely associated with urban living. Public transport and personal motor vehicles are major contributors to greenhouse gas emissions and air pollution in urban cities. Today's vehicles and on-road infrastructures are equipped with a large number of sophisticated sensory devices which are capable of monitoring and providing data pertaining to vehicle status, real-time traffic conditions, traffic incidents, and road crashes (Salim et al., 2008).

Urban commuters also carry sensors embedded in their mobile devices. Mobile devices and smart phones are now equipped with GPS, accelerometers, compasses (such as exhibited by the new iPhone 3GS). Clustering of population profiles, based on preferred transport mode, reveals trends that will be useful for government to formulate public policies for improvement of environmentally sustainable lifestyle (Salim 2010).

In addition, the recent rise of micro-blogging (Twitter), social networking (Facebook, Ning, Academia, and the like), and Web 2.0 powered with geo-tagging capabilities, if accessed from mobile devices, provides networks of emergence and convergence of spatial interactions and knowledge of and about our cities. This paper reviews some key components of designing a live social network in which the patterns of user interaction activities can be translated into useful information for establishing policies and programs to support sustainable urban living. Hence, the technological platform becomes the source of knowledge if viewed from government and city designer (architects, services design engineers, environmentalists, etc) perspectives. Hence, we term this kind of social network "knowledge network".

The Emerging Bits

Data Models

The first hurdle to data analysis is the lack of integration of various data sources and models in an urban environment. Modeling an urban environment requires a holistic view of the interconnectivity, given that a city is complex mesh of built environment models, land use and zoning, road networks, public transport networks, traffic models, energy and resource flows, and people's activities and movements. However, in operation, patterns of user interaction and activities are not well integrated with each other. Information is often not being exchanged and meaningful shared seamlessly between clusters within the "social network". Information is not being exchanged and shared between models. Fundamentally, an integrated and contextual urban information network needs to be established in order to inform the stakeholders of the city. SynCity is an urban energy simulation toolkit by Kierstead et al. (2009) that attempts to address the integration of these data models, however, this toolkit is still in

the prototype stage and does not address the difficulty in dealing with raw data and techniques to analyze data in order to generate information and knowledge about the city. The SynCity toolkit also implies the requirements for datasets that must be made available for the toolkit to work. This includes land use, transportation, energy use and flows, and socio-demographic data. However, in an urban context, such data are held by various stakeholders of the city, and releasing those data may be subject to privacy and security issues. In addressing this privacy concern, p-sensitive K-anonymity with generalization constraints algorithm can be used. This privacy issue, however, is not within the scope of this paper.

Targeted data collection presents different challenges. A small proto investigation into identifying energy usage in buildings across a university campus, immediately struck the difficulties that: (1) the power metering was neither consistently disaggregated by building, (2) nor by organizational unit (such as faculty or school), (3) the organizational units themselves were often distributed across parts of multiple buildings or even campuses. Moreover, the billing and purchasing arrangements with the power supplier were aggregated at institutional level. The age, quality and construction of the building stock across the campus was also extremely variable so in order to draw any useful information about the impact of construction or spatial organization on energy consumption and hence inference for retrofitting improvement, it would have been necessary to have data by building.

Although there is already a great deal of processed data, increasingly available in the public domain such as that provided by government agencies [1], these data are in disaggregated forms. Some data are provided in textual formats, others are tabulated, and some others in RSS feeds or XML. In order for data to be useful, they need to be given raw. Processed data impose particular uses of the data, whereas in an urban data model, the unseen emerging patterns need to be extracted from the pool of raw data. Collections of raw data from multiple domains need to be cleaned in order that data aggregation and management can be applied.

A major part of any urban research involving non-standardized data sources is the process of cleaning the data. Figure 1 shows the stages from data cleaning and integration to data summarization ready for analysis. In stage 1, raw data is being cleaned and integrated, and in stage 2, data summarization takes places. At the end of this stage, the data will be ready for analysis through various dimensions.

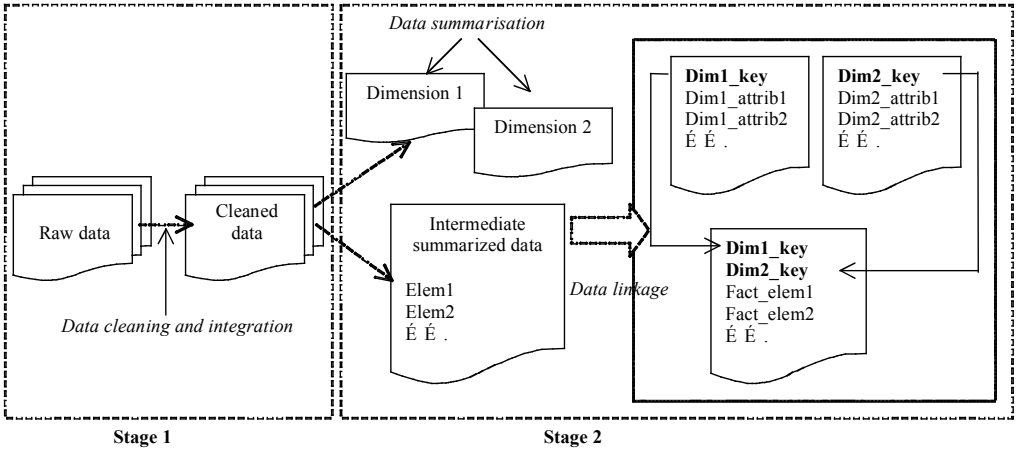

Figure 1
Overview of data preparation and processing

Raw data are collected from various sources, which are likely to have inconsistencies among these data sources, and therefore it is highly probable that some dirty data would exist. Dirty data includes missing data, wrong data, and inconsistent data representations (Kim et al., 2003). Dirty data can have a negative impact on the efficiency of data analysis, and hence, a thorough and careful cleaning of the data can save a lot of future effort and inconvenience during the subsequent steps of the data processing process. Generally, the data cleaning phase is the procedure of eliminating any anomaly which might exist in the raw data. In the cleaning process, two possible conflicts may occur among data sources: (i) schema-level conflicts, and (ii) data-level conflicts.

In the schema-level conflicts, also known as metadata-level conflicts, the raw data need to conform to some rules regarding their structure, elements, relationships, hierarchy, etc. Schema-level conflicts may exist in forms of naming conflicts, data type conflicts, or structural restriction conflicts. Data-level conflicts, on the other hand, appear when the information in the raw data is entered in different ways; some are due to incorrect data entry or inconsistent data formatting. Data-level conflicts may exist in a form of (i) data value conflict, (ii) data unit conflict, or (iii) data representation conflict. In order to solve these, various algorithms have been proposed (Rusu, Rahayu, and Taniar, 2004).

Crowdsourcing

Crowdsourcing is a relatively new word, only introduced back in 2006 by Jeff Howe (2006), who defines it as an act of outsourcing the tasks of developing new technology or application to the crowd through an open call. Howe (2008) presented four categories of crowdsourcing: crowd wisdom (or collective intelligence), crowd creation, crowd voting, and crowdfunding.

The examples of how the federal and state governments are interested in crowdsourcing through the provision of public data and use of Web2.0 technology are apparent worldwide. In Australia, the Government 2.0 Taskforce, was formed in response to the increasing interest of governments worldwide in the potential uses of public sector information and online engagement. This organization ran the competition MashupAustralia just over a month in late 2009 to elicit ideas about combining data sources for public communication, and raise awareness and public interest in open access to government data. They organized for the release of datasets from 15 agencies in addition to state and territory government on licensing terms permitting mashups. In web development a mashup is defined as a web page or application that combines data or functionality from two or more external sources to provide a new service. The best entries to this competition provided such services as comprehensive comparative suburb profiles including socio economic rating, education level, property information, family composition, crime safety etc for over 8000 suburbs; relating geographical area with the different levels of government associated with them, linking wildfire information to twitter to report and get up to date fire information. Each used from 5 to 20 different data sources in combination. This was a response to existing data sources, testing its usefulness and aiming to increase awareness, and ultimately levels of community participation. Following the open calls from the federal government, state governments are now following suit by holding app4nsw and App My State VIC, and expanding the competitions by also accepting mobile phone applications to be developed as well as mashups.

Crowdsourcing does not rely on data provision for it to work, as it can also rely entirely on the crowds to provide the data as long as the crowds can find the benefit and the engagement from the activity they are involved in. This is particularly demonstrated by *crowd democracy* activities, which could be part of crowd voting through the online platforms used solely to gather and present a community voice to the local, state, or federal governments. An example of this is Fix My Street [1], a site to report, view, or discuss local problems (such as graffiti, broken

road surface, pothole, or street lighting). Up to now, as this paper is written, there have been 879 reports in past week, 87,868 updates on reports, and 2,104 reported problems were fixed in the past month. The data reported on such applications can be used to analyze emerging patterns on the use of infrastructure and mobility of users of the city.

When mobile phones are loaded with applications that enable crowdsourced content to be submitted, the applications would evolve over time with developers only needing to create an intuitive placeholder for the crowdsourced content. *Crowd place-making* is spatial movement of the crowd that is voluntarily reported via mobile applications that can track their locations using GPS and geotagging and accept rich user inputs about their whereabouts. Foursquare [2] has become a popular social-networked place-making application and has been used widely given that the app for Blackberry and IPhone are available and it is connected to Facebook's status updates. Users of foursquare can "check-in" to places in the city, "shout" their status, and share tips about a place. If a place does not yet exist in the app, users can create new place, add information about the place, and add it to existing categories (i.e. restaurant, entertainment, education, transportation infrastructure, shopping, and so on). This type of applications is reliant on the crowd to populate the content and hence the existing application is capable of evolving from one use to another. Foursquare has also been used by shop and restaurant owners to promote their business by adding incentives and vouchers to those who are using foursquare to "check-in" into their locations. Hence, the application has evolved from a purely place-making tool to a marketing tool.

If we open a call for the crowd to collaborate with architects in designing a building, or larger, the city, is there a possibility for good designs to emerge? If there is an open platform for *crowddesigning* the city, how would we manage such a large scale participatory design?

To experiment with the question, James Willems Ransom from University of Buffalo, in collaboration with the first author, Przemek Jaworski from Foster + Partners, and Hugo Mulder from Arup, developed the TweetForm project in the SmartGeometry 2010 workshop using Twitter and parametric modeling [3]. TweetForm utilizes the power of Twitter as a social networking platform to inform a parametric model in Rhino. This enables crowdsourcing for collaborative form finding. Form finding activity is no longer restricted to individuals; instead, a much larger community can access the form, suggest modifications online, and preview the variations to occur in the form as suggested. The crowd needs to have access to a Twitter account, and post keywords to a specific hashtag associated to a particular Twitter account. The keywords, which are associated to particular geometrical variations in the model, include: lighter, heavier, taller, shorter, wider, thinner, rounder, squarer, smoother, boxier, twist, and straighten. Processing 3D application reads the Twitter stream and scans for the keywords. If keywords are found, data are sent via UDP directly to Grasshopper. Since the Rhino model is associated with the Grasshopper definition of generative behaviors as described by the keywords, the model gets updated accordingly whenever a keyword is received in the buffer. The 3D model was streamed online during the system testing, and was projected in one of the public spaces in the workshop venue in order to display the live updates from the Twitter stream (Figure 2).

Powered with Twitter and other social mediated means, we are seeing a wave of crowddesigned mobile and online applications, which are extendable by the crowds, only if they think they can benefit from these apps and if some forms of social engagements can be leveraged for networking purposes. If data provided by crowd can be analyzed, spatial movements and commuting experiences can be mapped and modeled against transportation network and built environment models. Group patterns (Wang, Lim, and Hwang, 2006) and movement patterns (Taniar and Goh, 2007) have been studied. Group patterns discover groups of people who are

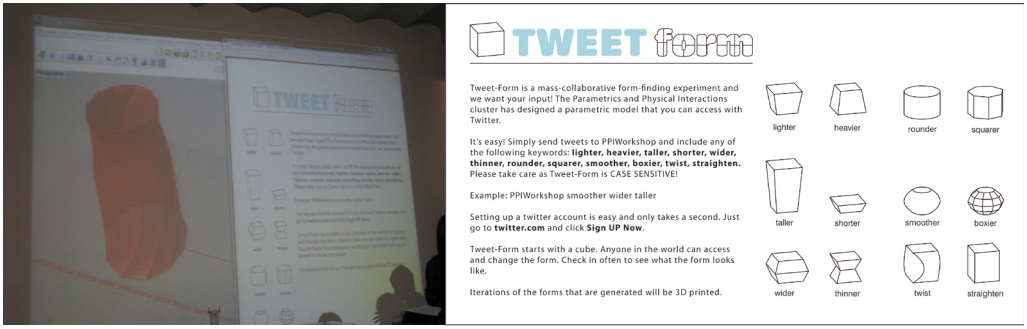

Figure 2
TweetForm

spatially together for a specified minimum duration threshold, whereas movement patterns examine groups of people who are moving together to certain directions and places. Group patterns and movement patterns can have a great impact to designing urban transportation, as they provide us with a tool to analyze people's movement. Both employ a data mining technique, in particular a modified version of association rules to discover these patterns.

Social Interventions

Persuasive Information Delivery via computer applications, when designed with the intention to modify user behaviors, has significant persuasive power (Fogg 2002). Fogg stated that a computer application can play the role of a persuasive social actor by rewarding the users with constructive feedback, simulating targeted user behavior, or giving social support (Fogg 2002).

The need for designing and developing effective persuasive strategies that target voluntary behavior change is crucial. The ubiquity of mobile devices provides effective means to deliver persuasive information to the widespread community. Using mobile technology, there is the potential for developing a persuasive mobile application that is capable of delivering real-time information about the city and at the same time monitoring personal contributions to the "green", safe, and sustainable measures of the city. An example of a persuasive mobile application is UbiGreen (Froehlich 2009), which encourages participants' engagement in using green transportation, such as carpooling and bike-sharing. UbiGreen collects users' behavioral data using semi-automated sensing and self-report questionnaires. UbiGreen uses trees or polar bear avatars to visualize the degree of 'eco-friendliness' by analyzing the user's weekly transportation choices. The study reports that the visual design employed is effective and sufficiently compelling to create awareness of green transportation options and persuasion toward greener lifestyle.

Aside from mobile applications, physical and responsive digital installations can be a persuasive aid to attract user interactions. This is exemplified in a summer project by Prateek Rungta, a Software Engineering student under the supervision of the first author. The assessment of surveys on workplace conditions, termed as Post Occupancy Evaluation (POE), was conducted in the recently built office tower in Monash University Caulfield campus, using a novel and non-traditional approach. POE study provides information that can be used either in modifying existing building systems or in subsequent building design to maximize productivity in the workplace by identifying previously successful combinations of variables under a building designer's control. These variables include, but are not limited to temperature, noise, ambience, space, cleanliness and lighting (Leaman 2004). The POE surveys are completed by buildings' occupants, since it is their comfort that the study seeks to evaluate.

The mashup idea was to replace the traditionally paper based POE surveys with a sensor powered system. A complete solution would be expected to handle most of the POE features aforementioned. Our prototype however, was built around just the temperature variable. Besides a temperature sensor, it used a proximity sensor to detect movement around the device. The mashup consisted of an Arduino circuit connected to a computer on which ran a controller program. The circuit components were the temperature sensor, the proximity sensor and two buttons, each accompanied by a feedback LED. The purpose of this setup was two-fold. First, it measured and reported the current indoor temperature. The controller program running on the computer would read, record and report these temperatures. It would mash this data with outdoor temperatures, fetched from the web (via the Bureau of Meteorology). The second task performed by the circuit was the automation of the POE. The infra-red proximity sensor would alert the circuit (and the controller) when someone approached the apparatus. The controller would then greet the person, report current indoor and outdoor conditions and ask if they were comfortable inside. The message plays via the computer's audio output. The user can respond by pressing one of the two buttons (labelled 'Yes' and 'No') connected to the Arduino circuit. These responses would also be read and recorded by the controller.

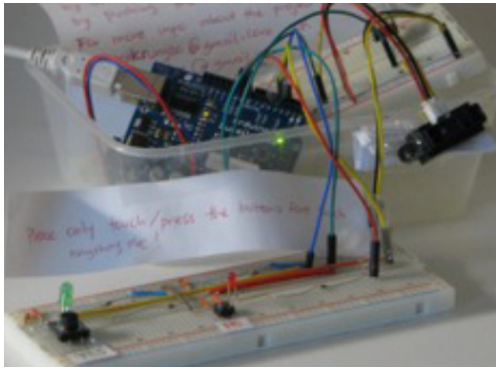

Figure 3
Automated POE physical tool

Thus the mashup automated the survey for temperature related comfort (Figure 3). The data collected could be mined to extract information useful for POE and other stakeholders.

If similar installations can be placed around the built environments for a crowdsourced POE, the data that are collected can be used to inform crowd-designing activities around the city. This will foster the growth of a knowledge network about the city and hence it will self-inform the city stakeholders of the real-time status of the city.

The Converging Bits

There are effectively three strands to this work. The first is building, augmenting or agglomerating existing data sources and rendering them in a clean, consistent and usable format. The second is to use data mining techniques to uncover hidden anomalies and relationships using combined data sources. And the third is to design prototype physical and social interventions within the city.

The first strand is clearly a very large ambition on the scale of the city and in the light of the project aims stated above. Within the scope of this project, we aim to firstly start to build a database of the data required to fulfill these aims, identify currently available sources and shortfalls, and at the same time investigate how it can be maintained "live" and current (with autonomous system design capability incorporate, it could be for a period of 3 to 4 years, after which major upgrading and redesign the entire system). As a testing cycle, we will build smaller demonstrators within this overall city data infrastructure model predicated on early accessible data and the strategic objectives of the project stakeholders.

The second strand can contribute to effective investment in retrofitting existing built environment and urban transportation. To increase our understanding of the recurring socio-economic behavioral patterns, a massive pool of data needs to be analyzed. Recurring patterns that appear in the data may lead to novel and useful information to

city stakeholders. Computational data analysis techniques are needed to discover such patterns. Data mining or knowledge discovery, which interconnects disciplines from machine learning, statistics, and databases, offers various techniques for data analysis, such as pattern discovery, clustering, classification, and time series. Pattern discovery in data mining exists in many forms, from the traditional association rules, to negative association rules and exception rules. Whilst the traditional association rules which are in a form of an event or an element followed by another event/element, negative association rules (Taniar et al, 2010) and exception rules (Taniar et al, 2008) capture different patterns, which are rare and sporadic. Hence, it is crucial to differentiate between noise/incidental and rare/sporadic but interesting rules. These patterns or rules are even more relevant in the built environment and urban transportation as anomalies and outliers are very important elements, especially when designing an effective environment. Furthermore, removing duplicate rules is a critical issue in order to make the rules more useful to decision makers (Ashrafi, Taniar, and Smith, 2007).

Knowledge discovery is also often a follow up of business intelligence, whereby intelligence techniques are drawn upon data warehouses. A data warehouse is required to assist decision makers and designers, as information in the data warehouse is integrated with various aggregators. Using a data warehouse, the management is able to drill down and roll up various levels of details of the required information. Data conflicts and cleaning is part of the data transformation process, which transforms the original data sources into an integrated data warehouse. Apart from drilling down and rolling up data aggregation along various levels, it is also possible to further perform pattern discovery on data warehouses (Tjioe and Taniar, 2005). One of the main differences between pattern discovery in data warehouses and in the original data sources is that patterns in data warehouses focus on aggregated data, which has a higher granularity than the atomic-level data in the original data sources.

Knowledge discovery and business intelligence (KDBI) techniques can be applied in order to build a contextual knowledge base of urban energy models that can potentially be used for retrofitting existing buildings, informing new developments, informing new land use, traffic and public transport optimization, supporting policy revision for carbon reduction, energy supply and demand matching, and modeling targeted marketing approaches for sustainable lifestyles. KDBI (Figure 4), in a nutshell, is a process of transformation of data to information through data management (such as Database Management or DBMS), which then generates information that need to be translated to knowledge and business intelligence through application of data mining techniques and optimization (Michalewicz 2007). At this juncture, we wish to reiterate that the goal of this paper is not to articulate the roles of each specialist, it is to propose an automated system, of course taking input from domain specific expert knowledge into such design.

The third strand is synthetic and also seen as exemplary: to design prototypical physical and social interventions within the city, which inform, reward and modify behavior. With regard to this last endeavour, we will take advantage of mobile technology to tailor information and communication to personal profiles. Thus the "greenness" of an individual's choices and actions will be moderated by their particular context. A parent of young children living in an outer

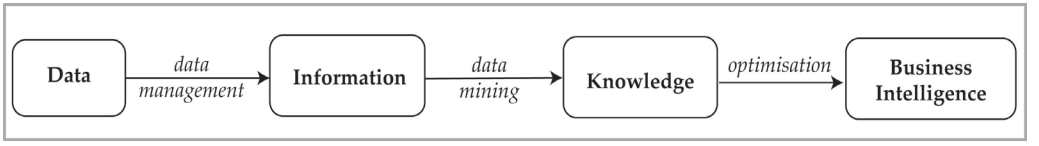

Figure 4
Transformation of data to business intelligence

suburb may have more limited opportunities to vary their choice of transport to reduce carbon emissions, for instance, than a student living in the city centre. Thus the reference to lifestyles refers to customized interaction. Crowd-designing such applications in the context of urban interactions can be used to inform the first and second strand, which eventually is a convergence of a top-down and bottom-up approaches to setting up a live knowledge network for urban sustainability.

References

Ashrafi, MZ, Taniar, D, and Smith, KA 2007, 'Redundant association rules reduction techniques', *International Journal of Business Intelligence and Data Mining*, 2(1), 29-63.

Fogg, BJ 2002, *Persuasive Technology: Using Computers to Change What We Think and Do (Interactive Technologies)*, Morgan Kauffman.

Froehlich, J, Consolvo, S, Dillahunt, T, Harrison, B, Klasnja, P, Mankoff, J and Landay, J: 2009, 'UbiGreen: Investigating a Mobile Tool for Tracking and Supporting Green Transportation Habits'. *Proceedings of CHI2009*. Boston, MA, USA.

Howe, J 2006, 'The Rise of Crowdsourcing', *Wired*, Issue 14, no. 06, June 2006.

Howe, J 2008, *Crowdsourcing: Why the Power of the Crowd is Driving the Future of Business*, 1st ed., Crown Publishing Group, NY, USA.

Keirstead, J, Samsatli, N, and Shah, N 2009, 'SynCity: an integrated tool kit for urban energy systems modelling'. In *Proceedings of the 5th Urban Research Symposium*.

Kim, W, Choi, B-J, Hong, E-K, Kim, S-K, and Lee, D 2003, 'A Taxonomy of Dirty Data', *Data Mining and Knowledge Discovery*, 7 (1), 81-99.

Leaman, A 2004, 'Outside the comfort zone: building human and basic needs', *Human Gives Journal*, 11 (2).

Michalewicz, Z, Schmidt, M, Michalewicz, M, and Chiriac, C 2007, *Adaptive Business Intelligence* , Springer-Verlag, Berlin Heidelberg.

Rusu, LI, Rahayu, W, and Taniar, D (2004), 'On Data Cleaning In Building XML Data Warehouses', *Proceedings of the Sixth International Conference on Information Integration and Web-based Application Services (iiWAS'04)*, Austrian Computer Society.

Salim, FD, Cai, L, Indrawan, M & Loke, SW 2008, 'Road intersections as pervasive computing environments: Towards a multiagent real-time collision warning system', *Sixth Annual IEEE International Conference on Pervasive Computing and Communications*. Hong Kong, China, IEEE Computer Society.

Salim, FD 2010, 'Towards adaptive mobile mashups: opportunities for designing effective persuasive technology on the road', *Proceedings of the 24th International Conference on Advanced Information Networking and Applications Workshops (AINAW'10)*, IEEE, US.

Taniar, D, and Goh, J 2007, 'On Mining Movement Pattern from Mobile Users', *International Journal of Distributed Sensor Networks*, 3(1), 69-86.

Taniar, D., Rahayu, W., Lee, VCS, and Daly, O 2008, 'Exception rules in association rule mining', *Applied Mathematics and Computation*, 205(2), 735-750.

Taniar, D., Rahayu, W., Lee, VCS, and Daly, O 2010, 'Mining Hierarchical Negative Association Rules', *International Journal of Computer Systems: Science and Engineering*, 23(2).

Tjioe, HC, and Taniar, D 2005, 'Mining Association Rules in Data Warehouses', *International Journal of Data Warehousing and Mining*, 1(3), 28-62

Wang, Y, Lim, E-P, and Hwang, S-Y, 'Efficient mining of group patterns from user movement data', *Data and Knowledge Engineering*, 57(3): 240-282 (2006)

[1] http://www.fixmystreet.com/

[2] http://foursquare.com/

[3] http://ubimash.com/