HELSINKI UNIVERSITY OF TECHNOLOGY
Department of Electrical and Communications Engineering

POLITÉCNICA

Escuela Técnica Superior de Ingenieros de Telecomunicación
Universidad Politécnica de Madrid

Borja Jiménez Salmerón

MODELING OF MOBILE END-USER CONTEXT

Thesis submitted in partial fulfillment of the requirements
for the degree of Master of Science in Technology

Helsinki, May 20th, 2008

Supervisor	Heikki Hämmäinen Professor, Networking Business
Instructor	Hannu Verkasalo Lic.Sc. (Tech)

HELSINKI UNIVERSITY OF TECHNOLOGY

Abstract of the Master's Thesis

Author:	Borja Jiménez		
Name of the Thesis:	Modeling of Mobile End-User Context		
	20.5.2008	**Number of pages:**	98+7
Department:	Department of Communication and Networking (Faculty of Electronics, Communications and Automation)		
Professorship:	S-38 Networking Technology		
Supervisor:	Prof. Heikki Hämmäinen		
Instructor:	Hannu Verkasalo Lic.Sc. (Tech.)		

Emerging mobile services have spawned new revenue sources, such as messaging, Internet browsing and multimedia. One of the business opportunities that mobile industry has not yet fully exploited is the contextual status of end-users. Context-aware systems are gaining importance in telecommunications since the applications are numerous and have relevance from the industrial (e.g. in aspects involving user segmentation) and academic (e.g. analysis of mobile service adoption dynamics) points of view.

This thesis first presents a theoretical discussion: evolution of telecommunications services, prior studies in context-aware systems and other concepts such as data mining techniques and network theory, and network visualization. The second part focuses on the development of a context detection algorithm. This algorithm extracts contextual information from data logs containing cell-id transitions. It follows two steps in context detection: first a clustering process where physically close cells are grouped into clusters and second the context detection for every one of those clusters by using time-based assumptions. The thesis uses a handset-based tool in collecting data logs. The strength and accuracy of the algorithm are tested through analysis of the output files. Finally, a study of real data (from the Finnish market) is carried out in order to deliver results. Through this analysis, the thesis focuses on the service usage perspective. The driving question is how and where the end-users spend their time with the handsets.

Context is not only about location but also about the physical status and social settings of end-users. Context detection provides a new dimension for example in service usage analysis or modeling of service adoption. The results show that e.g. most of the usage of WLAN takes place at "home" and applications such as "Navigation and Maps" or "Browsing" are used "on the move" context. Intensity graphs prove that e.g. "Home" is not the most active context despite being the most frequent one and the intensity of usage abroad is most active in "Multimedia" and "Messaging" applications. On the other hand, there is a significant business opportunity in applications that automatically identify the context of mobile users (all their movements along the day). Targeted marketing and handset-based contextual adaptation are some of the examples of possible applications. At last, the thesis confirms that network visualization tools are useful in the process of context modeling, not only for testing the results but also to help in the detection by using all their functionalities (as e.g. clustering). Some examples using one specific tool will be provided at the end.

Keywords: Context, Modeling, Algorithm, Visualization

ESCUELA TÉCNICA SUPERIOR DE INGENIEROS DE TELECOMUNICACIÓN
UNIVERSIDAD POLITÉCNICA DE MADRID

Anteproyecto

Nombre:	Borja Jiménez	
Título del proyecto:	Modelado del contexto de usuarios móbiles	
Contexto,	20.05.2008	**Número de Páginas**: 98+7
Departamento:	Departamento de Comunicaciones y Redes (Facultad de Electrónica, Comunicaciones y Automática)	
Cátedra:	S-38 Tecnología de Redes	
Supervisor:	Prof. Heikki Hämmäinen	
Instructor:	Lic. Hannu Verkasalo	

La aparición de múltiples servicios móviles ha generado nuevas e importantes fuentes de beneficio tales como la navegación por Internet o los servicios multimedia. Una de las oportunidades de negocio abiertas, no explotada aún por la industria móvil, es el estatus de contexto de usuarios móviles. Los sistemas de reconocimiento de contexto han ganado importancia en el mercado de las telecomunicaciones con el paso de los años, aportando gran valor añadido tanto para el entorno privado (en aspectos relacionados con segmentación de usuarios) como para el académico (análisis de adopción de servicios).

En primer lugar se presenta la base teórica: evolución de los servicios móviles, estudios iniciales sobre sistemas de detección y otros conceptos tales como técnicas de minería de datos, teoría de redes o herramientas de visualización de redes. La segunda parte se centra en el diseño y desarrollo de un algoritmo de detección de contexto. Dicho algoritmo es capaz de extraer información de usuario a partir de ficheros con información acerca de sus transiciones entre células de cobertura. La detección del contexto es un proceso que se divide en dos pasos: en primer lugar la “extracción de clústers” donde las células próximas físicamente son agrupadas en clústers y, en segundo lugar, la “detección de contexto” para cada uno de los clústers hallados usando el tiempo como parámetro clave. La eficacia y eficiencia del algoritmo son validadas analizando los ficheros resultantes. Por último, se elabora un estudio usando datos reales (mercado móvil finés) con objeto de provisión de resultados.

El contexto no se refiere sólo a la localización sino que también hace referencia al estatus físico y al entorno social del usuario. La detección del contexto proporciona una nueva dimensión en análisis de uso del dispositivo móvil o en aspectos relacionados con la adopción de nuevos servicios o tecnologías. Los resultados muestran, por ejemplo, que la mayoría de uso de WLAN tiene lugar en entorno “doméstico” y que aplicaciones tales como “Navegación” o “Mapas” son usados básicamente en contexto “en movimiento”. El uso de gráficos de intensidad de uso muestra como “en movimiento” u “oficina” son contextos más activos que el contexto “doméstico” en la gran mayoría de servicios y aplicaciones. En último lugar, la tésis se centra en el uso de herramientas software de visualización de redes para el análisis de resultados y como soporte en la detección de contextos usando sus múltiples funcionalidades.

Palabras clave: Contexto, Modelado, Algoritmo, Visualización

Acknowledgements

First of all, I would like to thank both Heikki Hämmäinen and Hannu Verkasalo for the great opportunity they gave me in the moment they accepted me to carry out my thesis for this laboratory and for all the time, ideas and dedication they invested on my success. I wish also to show my gratitude to all the team members that made me feel at my home university all the time.

I want to remind here to my family how lucky I am for having their support in every decision, their help in every bad moment and all their love. I never walk alone. This moment is also dedicated to my grandparents and to my aunt. Without all of them nothing would have been possible.

I could not forget to thank all the people that shared this experience with me, in Helsinki and in Madrid, making it unique in every sense.

And of course, last words for all the "*good friends we have, good friends we've lost, along the way…*"

Espoo, Finland, May 2008

Borja Jiménez

Table of Contents

List of Figures

List of Tables

1 Introduction

1.1 Motivation

The mobile industry continues expanding. Since the early 90's there has been a peak of mobile services usage where the concept of Triple-Play (marketing term for the provisioning of the two broadband services, i.e. high speed internet access and television and one narrowband service, i.e. telephone, over a single broadband connection) is intimately bound up with the concept of mobility (Triple-Play plus mobility is known as well as Quadruple-Play, see Bauer 2005). The reason of this peak is easy to understand: the mobile telecommunications market has experienced an unprecedented growth, without similar examples along the recent history of new technologies. The access to every service while moving is one of the principal needs for the users and, therefore, it is nowadays one of the most promising and profitable business for network operators and applications developers.

There are several entities involved in the evolution of the industry. On the one hand network operators have seen how a big new business appeared. On the other hand the users have adopted a new technology with wide penetration (that nowadays has finally stopped due to almost everybody has a mobile phone and, in many cases, two or even more). Application developers, Internet service providers and content producers should not be forgotten either. The regulators are also involved in the growth with a very important role (telecommunications law and policy framework).

The telecommunications market was earlier considered a “walled garden” because operators control the whole portfolio of services. But this situation has changed fast. Smartphones and Internet services in mobile phones are breaking the walls of the garden down forcing the operators to adapt to the new situation (see Chau 2007 or Best 2006).

In order to better serve the market, one of the most important priorities for big companies interested in segmentation is the increase of collected data regarding to end-users behaviour. A better knowledge of the user likings, habits or behavioural modeling becomes possible with the new data. The aim of segmentation is to detect user habits and preferences in order to create groups of users. One of the alternatives that this matter leaves open is the modeling of mobile end-user context. To be able to know the context of every concrete user (and aggregate results) means that, from the company's point of view, they have better knowledge of their users, they are able to know what applications do users use more often or where the user buys (physically or via internet). Even more important, they can know where, when and how users utilize their services and how they interact among them.

Context-aware systems are gaining importance in telecommunications and a big number of context studies have been carried out in the past years. Context is much more than the mere location; it is something that relates to all situational factors around the user in a particular use case (Kaasinen 2003). If the proper data is collected and analyzed, service providers can extract very valuable knowledge related to the e.g. the context of usage. Therefore, context is not only where the user is in a particular moment, but also what he does with the handset, with whom, how, etc... This new source of knowledge helps the operators to know more about their users and let them to think about new applications to launch, how to focus advertising campaigns or to know where they should place more resources according to user habits.

Contextual modeling is a new opportunity for academic contributions as well. Users take their handsets with them along the day. Mobile devices report daily usage actions and, this way, operators and researchers can know e.g. if the end-user has Wi-Fi at home (once they are able to determine different contexts, service usage analysis under a contextual perspective can be carried out), what music artists they listen to (another possibility for focused and personal advertising), the websites they visit more often or how they connect with their friends, creating social networks that open another world of knowledge about user behaviour. For the academia contextual data adds a new dimension in modeling network usage, adoption processes or behaviour of end-users.

A handset-based end-user research method with a pioneering data collection platform facilitates new research topic. Whereas earlier hardware limitations have challenged the attempts to acquire subscriber-level usage data, the developed handset-based technology supports new kinds of end-user research (see Verkasalo 2007a and Verkasalo 2007c). The interest of network operators and service providers in the use of this technology to explore their users, for obvious reasons, is continuously increasing (Verkasalo 2005).

1.2 Problem

Most mobile users move actively during an average day although some people move more than others. Cell-ids can be used as an analogy to geographical location and further to context of mobile users**. Based on accurate handset-based data logs, how to transform data on cell-id transitions to information of location and context?** In particular, how to visualize this and how to build an algorithm that automatically processes huge amounts of cell-id location data and cluster cell-ids into geographical locations?.

How to detect different contexts, how to represent in a picture the user's location, how to generalize the results, how to handle the raw data, how to simplify this raw data without losing relevant information are just some problems this thesis have to solve and are important parts of this study.

The research questions of this thesis include:

1. How to detect different contexts?
2. How to visually present the user's location?
3. How to automate contextual data processing?

This thesis evaluates a range of software tools regarding network visualization. These tools offer algorithms to present data and results graphically. It is part of this research to analyze the state of the art and find an appropriate tool from all to be used in the process of context modeling.

With the help of the selected tools and the knowledge provided prior this thesis aims to answer the anderlying research question: **How can various data mining approaches be used in extracting contextual information from handset-based data?**

1.3 Objectives

The key objective of this research is to model the context and location of mobile users from the raw data available (through the clustering of cell-ids) and to visualize the information graphically by using specific software tools. For this purpose, an algorithm to automate the process for large amounts of data is developed and used with a real Finnish smartphone panel data provided by Nokia. Once the location and context can be extracted from the data this thesis focuses on the future applications.

The aim of the thesis is to detect the possible contexts for every single user rather than to identify the exact location. The more contexts identified the better accuracy and proximity to reality of the representation done (e.g. if instead detecting only “on the move” context, the algorithm detects “driving” or “walking” the valued added is higher).

1.4 Scope

Modeling of mobile end-user context is a new technological opportunity. Thus, not much investigation has been done on contextual issues (see Verkasalo 2007b, Esbjörnsson & Weilenmann 2005, Lee et al., 2005 or Kaasinen 2003).

Inherent limitations exist when working with user data samples e.g. in the generalization of the results. The data collection used is biased (provides the information of early-adopters having smartphones) and small (from 500 to 2000 different users).

Furthermore, empirical data collected straight from devices need to be processed to reduce redundant information and complexity. This implies an inevitable loss of information.

At some point of the analysis, decision making based on subjective criteria is necessary to extract conclusions (e.g. percentage limits). This will be another important limitation of the approach taken. For instance, it is necessary to reduce the number of target locations (the id-cells will be marked as home, office, on the move and abroad).

Limitations exist in the presentation of results as well. Due to software limitations (network visualization programs usually present drawbacks while working with networks of more than 100 nodes) the data complexity reduction explained before becomes necessary.

1.5 Methods

A literature study is carried out to study context-aware systems, data mining approaches, and network theory. This study is based on books and research papers.

The development of an algorithm for processing large datasets containing end-user information (transitions between base stations) is an important part of this thesis. In this algorithm, the basic steps are the data processing (via clustering and time-based calculus) and the use of some heuristics (for the context identification of every cluster).

Testing of the algorithm is based on a set of trials, controlled simulations and a sensitivity study explained in the Appendix B..

Service usage is studied with SPSS by using the results of the context modeling algorithm.

1.6 Structure of the Thesis

First, a literature review of technologies used in the thesis is presented (mobile service evolution, context theory, data mining and network theory).

Secondly, the algorithm developed to model the context is explained theoretically. The basic assumptions, rules and the pseudo code of this software are described in detail, especially the parts more related to the clustering process and context detection.

Next, the application of the algorithm is demonstrated using real market data from Finland 2007. The data set, the algorithm's configuration (established after a sensitivity analysis) and the results of a contextual mobile service usage study are described.

The main findings of this research are presented in the final chapter, as well as some examples of possible exploitations, limitations and future research.

2 Background

A background for mobile service market is presented firstly (under the perspective of the new Internet services and the spread of smartphones use). In the second section, the handset-based collection of data used is introduced. Next a review of prior studies focused on the definition of context is discussed to understand the concept of context modeling. The most relevant aspects of network theory are studied as well since they will be used in the process of context visualization. Due to the data processing required, an overview of the data mining fundamentals is introduced in the following section. The utility of the network visualization software and some of the tools available are examined in the last section.

2.1 Mobile Service Market

Mobile industry is suffering a transformation. This creates new opportunities and challenges for infrastructure, service providers, end-users, and for the future direction of the industry. The established value chain is being deconstructed at the same time as some of the classic and tested business models become obsolete, what has forced the players to redefine their strategies and market positions. One of the milestones of telecommunications industry evolution has been the deregulation of telephone services (that started in the US in the eighties with the AT&T division, followed soon by the UK and Japan), On the other hand, the liberalisation of the telecommunications regulatory regime has allowed new licensed entrants to compete against the incumbent. But despite the big changes forced by deregulation and liberalization, the benefits have been significant: an improved service provision and quality, price reductions, service innovations and modernization. (Li et al., 2002)

An important key factor in telecommunications evolution has been the development of the Internet technologies. With the advent of the Internet and the plurality and variety of new applications brought in, the need for new and more advanced services in cellular phones has increased rapidly. However, new technologies irruption has not result in the apparition of new mobile services. This has been motivated for the intrinsic differences between the Internet and the mobile telecommunications systems that have made the adoption more difficult than expected. Digitalization is breaking telecommunications and computer networking barriers making possible to use same technologies in both fields. Unfortunately, convergence is proceeding slower than expected and it has been proven the impossibility of moving applications from one field to another in many cases. The reason is that although technologies in use are rather similar there exist essential differences in architecture and concepts. (Jorstad et al., 2004)

Less than a decade ago, end-users were served by proprietary networks where the phone network carried voice, the local area network carried data, and the broadcast network carried video. Each of these networks could be considered as a vertical pipe (i.e. closed systems). Because of the Internet, the vertical disintegration of these pipes has been enabled, allowing new entrants to enter the telecommunications market. Before the appearing of the Internet, the liberalization and the deregulation, the telecommunications industry was divided into three layers: equipment, network and services. Nowadays it is possible to identify six different layers: equipment and software, network, connectivity, navigation and middleware, applications (including contents) and customers. This change has motivated the apparition of new players, new opportunities and the openness of the industry. (Li et al., 2002)

Among all the new business that the Internet has made possible, special attention should be paid to those focused in the interconnection of users (where the value added for the incorporation of a new member in the network does not increase in a linear way). People have used the social network metaphor for over a century to connote complex sets of relationships between members of social systems at all scales, from interpersonal to international (J. A. Barnes started to generalize the term and concept in 1954). In addition, it has been used to explain links between people inside the companies (e.g. information flow or company culture) and communities (e.g. friend, student, social or cultural). The popularity of Social Networking online has experienced an explosion and websites like

MySpace, Facebook, YouTube or Flickr (very good examples of social networking applications) are expected to exceed 230 million of active memberships by the end of 2007. The social networking industry and telecommunications and media industries have started looking at how such services and user-generated content can be commercialized on mobile phones. This is the reason why companies like Radar, Zannel or Jaiku are working to allow people to create networks of friends that stay in touch through their mobile phones. (WinterBottom 2007)

Similarly, Internet heavyweights do not want to let this opportunity slip either and they have started to form alliances with operators and developers to offer their applications on the mobile phone market (e.g. BlackBerry devices and Facebook, Virgin Mobile USA and Facebook, or Vodafone U.K. and EBay, MySpace, GoogleMaps and YouTube). The reasons of this interest are very obvious: production costs associated with user-generated contents are very low and the associated increase of data traffic (the only way for operators to increase the ARPU nowadays). Because of the horizontal extension of the mobile services industry, the future of this field is promising for operators, handset developers, application providers and researchers. (WinterBottom 2007)

Context-aware computer applications were the first attempt to identify end-user contexts for different purposes, followed soon by the smartphones. Because of the vague idea of context and its generic definition, the concept of context can be interpreted in too many ways, depending on the specific application. Sensors-based context-awareness for adaptive Personal Digital Assistant user interfaces (see Schmidt et al., 1998), location-aware information delivery systems (see Chen et al., 2001), context-aware nomadic information systems providing adaptation (see Gross et al., 2001), adaptive on-device location recognition systems (see Laasonen et al., 2004), adaptive mobile phone applications (see Esbjörnsson et al., 2005, Raento et al., 2005 or Padovitz et al., 2005), mobile context-aware tour guide applications (see Long et al., 1996) or context-aware computing applications regarding location (see Schilit et al., 1994) are just some examples of the research done in this field.

This thesis focuses on one concrete application: the modeling of mobile end-users context. Definitions of context and related work will be presented in the following sections.

2.2 Handset-Based End-User Research Method

Based on handset-based data logs, this research explains a way to transform the cell-id transitions contained in these data into context information. The handset-based data of mobile end-users used in the thesis has been presented in several papers (e.g. Verkasalo 2005, Verkasalo 2007a or Verkasalo & Hämmäinen 2006).

The handset-based data collection has overcome many problems in the study of mobile end-user's usage regarding subjectivity. First of all, user interaction and his/her perception disappear because the monitoring system acquires the data straight from the devices. Secondly, the accuracy of the data obtained without human interaction is higher: a handset monitoring software can measure the usage frequencies, durations and volumes of all terminal features and applications (Kivi 2007). However, the data present some limitations explained in the first chapter that have severe influence on the results (they are a sample of a specific - early-adopters of smartphones - limited number of users). Besides this, the processes of collecting and recognizing locations from the cell id transition data logs are challenging for several reasons e.g. cellular network dimensions (cells can be very large), overlapping areas covered by different base stations, uncertainty about the physical topology of the cells for academic researchers (only network operators know the exact location of their base stations), cell changes do not occur always in the same actual location when moving (due to the existing lag in changing cells to minimize network traffic) or cell changes not as a result of changes in location but as a product of an interference or a power loose (although they talk about GSM networks and the data referred here have been obtained from WCDMA networks, see Laasonen et al., 2004).

One of the objectives of this thesis is to design and implement a general algorithm that processes inputted cellular network logs of mobile phones. As it is explained in the following chapters, the most relevant aspects of the algorithm are those related to time (timestamps) and, of course, to location (base stations where the users are in a certain moment of time). Thanks to the handset-based data collection method mentioned above, the handsets are able to register every activity in several types of files. Finally, the research platform sends daily reports to the data bases where the raw data of end-users are stored. This material can be used not only to test the context detection algorithms developed, but also to interrelate topics as segmentation and user trends, mobile business market studies, social networks or new technologies impact.

2.3 Context Modeling

Several studies related to context and context-aware services conclude that a definition of the concept of context is something extremely difficult and it varies depending on the particular situation and the aim of the study. There is no consensus in the scientific community about several questions related to the term either although attempts to create a standardized definition of use-context have been made (ISO 13407, 1999). This way it is not yet possible to determine without controversy whether the context is e.g. static or dynamic, external or internal, a set of information or processes, or finally, a simple set of phenomenon or an organized network. (Bazire & Brézillon, 2005)

Kaasinen goes further asserting that contexts of use in mobile environments are something that vary a lot and may even be continuously changing during use (see Kaasinen 2003). The location can be considered as an element of the context (and it will be measured more or less accurately depending on the positioning system in use or, in the case of this study, on the accuracy of the handset-based data available).

The Encyclopaedia Britannica (2007) defines the context as “the interrelated conditions in which something exists or occurs (environment setting)”. For the interest of this study, it could fit better the definition already used in the first chapter: “Context is much more than the mere location; it is something that relates to all situational factors around the user in a particular use case” (Kaasinen 2003). But, there is no full definition of context that can be enough for the purposes of this study. All the “situational factors” could be personal, physical and social, factors that can be highly influenced for the environment. Mobile phones and services can be used in a wide variety of contexts. Under this and the applications integrated perspective, it is probably more accurate to refer to the term context as “use context”, since this study is focused on detecting e.g. not only where the mobile user use the mobile phone, but also what he/she is exactly doing with it.

Many prior studies have tried to define and categorize the main elements of the context. As it has already been said, the difficulty on defining context can be moved on to the attempt of classifying the parameters in which context can be classified. For this reason, for every deep study found, a new way to describe context appeared. It is going to be useful for the thesis to present some of these context descriptions.

Schilit et al., (1994) inspired many later studies with their research about context-aware computing. In their paper, they assert that context includes more parameters than location and they name the most important aspects of context: where you are, who you are with and what resources are nearby.

Based on these ideas, Gross & Specht (2001) established a basic classification for detecting context types in nomadic information systems (see below). The idea of using “Environment/Activity” as a parameter is very practical. It describes the artefacts and the physical location of the current situation.

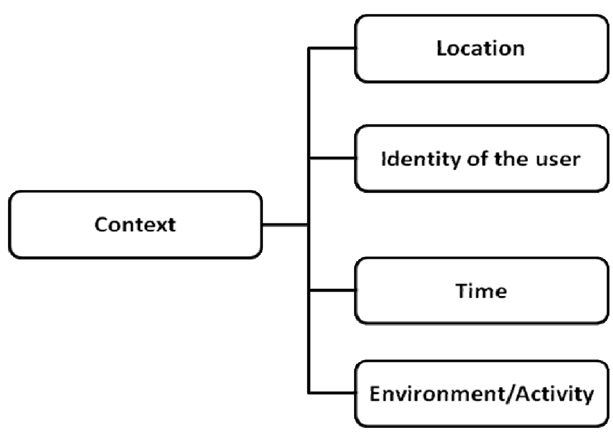


Figure 1 - Main dimensions of context (adapted from Gross et al., 2005)

Schmidt et al., (1999), define context by separating again the ideas of “human factor” and “physical environment”.

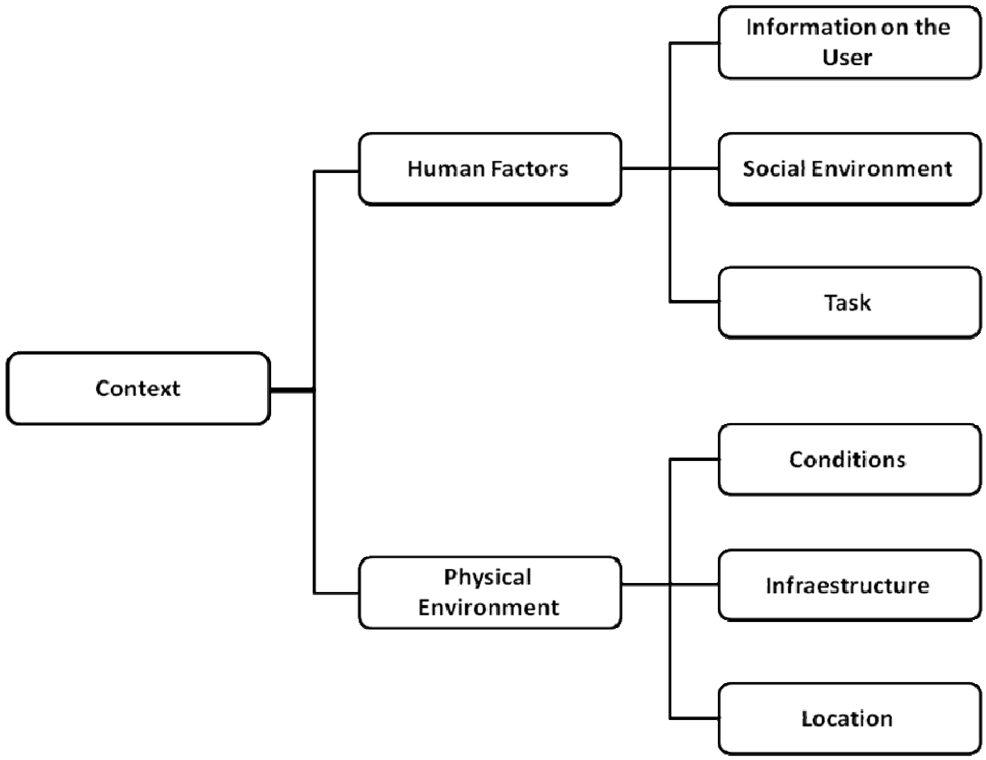


Figure 2 - Context feature space (adapted from Schmidt et al., 2005)

Information on the user refers to habits and emotional state, social environment provides information about co-location of others, social interaction, group dynamics and the user’s tasks talk about tasks in a working context. On the other hand, in the physical environment branch, conditions refer to physical conditions as pressure, noise or light among others, infrastructure talks about surrounding resources and location means the position in space, in the company, in the hierarchy...

In a more practical study regarding context-aware mobile computing, Chen et al. (2001), presents a different classification of contexts based on the ideas of Schilit, as shown in the next figure:

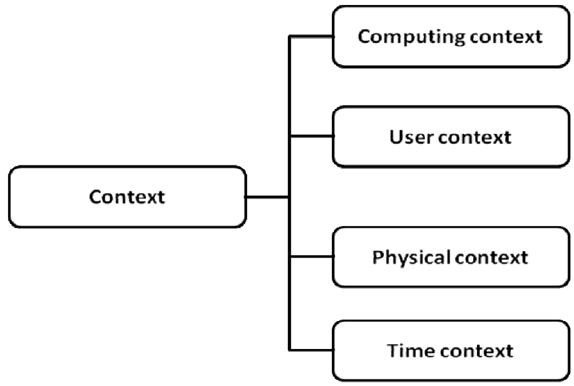


Figure 3 - Schilit definition of context (adapted from Chen et al., 2001)

"Computing context" refers to network connectivity, communication costs and bandwidth, and nearby resources as printers or workstations. "User context" identifies user's profile, location, people nearby or social situation. "Physical context" means noise, light or temperature.

Despite these ideas are vague and abstract for practical issues in most of the cases, all the previous research inspired further and more accurate studies on use context detection. Lee et al. (2005) go further presenting the classification of context based on the studies commented above. Their definition is shown in the following figure:

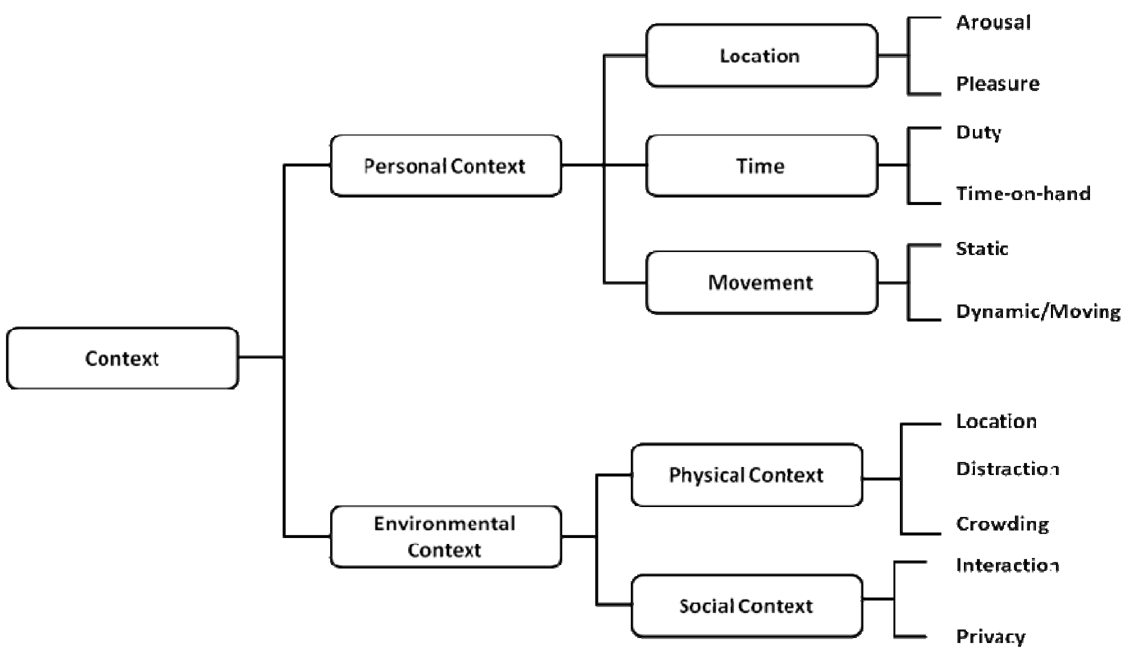


Figure 4 - Conceptual framework for context of usage (adapted from Lee et al., 2005)

Although another different classification focused on the purposes of this thesis is presented and explained in further chapters, earlier research has proven that context can be understood as a group of parameters that includes e.g. location, timestamp or current use (application or service used in the moment of the report) among others.

2.4 Network Theory

As an important aspect linked not only to visualizations but also to the interpretation of the handset-based data used, some definitions related to the network theory must be provided at this point.

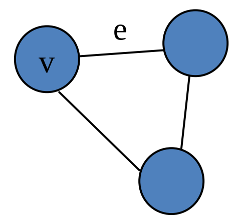
A network consists of a graph and additional information on the vertices and the union lines of the graph, where a graph is a set of vertices (the smallest unit in a network) and a set of lines between pairs of vertices. A vertex represents the actors or other abstractions of the real life (like e.g. computers in computer networks, people in social networks or processes in flow graphs) and the lines between them represent different kind of relations that link these vertices considering the logic of the specific network used. (De Nooy et al., 2005)

The additional information mentioned in the definition of network can be included in the vertices by using sizes (even colours or labels) and in the links between vertices. These connections can be directed (i.e. arcs) or undirected (i.e. edges or bidirectional arcs). The links can be weighted as well, giving more information by the simple use of a number in the connection. These numbers can have different meanings depending on the network type (e.g. traffic rate exchanged between two nodes that represent a computer network).

Networks theory and modeling can be used in a vast set of problems where actors and relations among these actors have the valuable information. Networks are used to represent different kind of structures in disciplines like telecommunications (e.g. computer networks, information networks or traffic networks), biology (e.g. molecular networks) and even sociology (social networks are gaining importance in the scientific community and in the private market due to its relation to the new business orientation and the network externalities: the more users the more profit, see Liebowitz et al., 1996). Wasserman et al. (1994) suggest the application of the concepts of social networks (considered as one of the future killer applications, see Karp 2008) in fields like marketing, economics, and industrial engineering. But, as already said, networks are used in many other disciplines as well. Recent papers like e.g. Rohlf et al., (2008) study the network structure of artificial genome by analyzing the statistical properties of artificial genome networks. On the other hand, since the very beginning of Telecommunications, the use of network representations to describe computer networks or visualize router connections that helped to analyze structure, packet flow or traffic and congestion (see Jahn et al., 2005) has been an important part of the theory studied. As example, see additional papers related to the evolution of IP networks (Bernardos et al., 2005), Wireless Networks Security (Xiao et al., 2007) or business value in social networking applications (IDC 2007).

The concept of graph, so often used to describe what a network is, comes from mathematics (graph theory specifically) although it has been always related to computer science as well. Because of this, the underlying mathematical theories and calculus associated to network modeling cannot be obviated. Some basic but necessary definitions are provided in this section. Using graph theory, a graph or undirected graph *G* can be defined as an ordered pair (see Caldwell 1995),

$$G := (V, E)$$



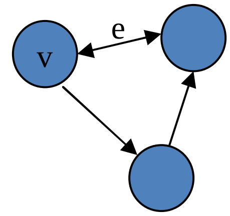
Where,

- V is a set of elements called vertices or nodes.
- E is a set of pairs (unordered) of distinct vertices, called edges or lines.

V and E are usually finite sets. The order of the graph is defined as the number of vertices and the degree of a vertex is defined as the number of other vertices connected to it by edges.

In the same way, a directed graph or digraph G is an ordered pair,

$$G := (V, A)$$



Where,

- V is a set of elements called vertices or nodes.
- A is a set of ordered pairs of vertices, called directed edges, arcs or arrows.

In directed graphs, the arc a is defined as,

$$a = (x, y)$$

Where x is the head and y is the tail.

From these two basic definitions, an important set of more complex graphs can be derived. These graphs can be used in several disciplines to solve big range of problems. They can be weighted and bipartite graphs deserve special attention for the interests of this study. A graph is weighted if a positive number (weight) is assigned to each edge. The weights might represent, e.g. costs, lengths, capacities or rate of traffic exchange. Weighted graphs are widely used in telecommunications (see Balakrishnan, 1997). On the other hand a bipartite graph is a graph with two kinds of vertices e.g. *W* and *X*, in which there are only edges between two vertices of different kind (Peltömaki 2008). The interest of bipartite networks lies in its possible applications to social networks and, therefore, to segmentation.

A very important part of the network theory is the formation of classes, groups, families or partitions of nodes. The understanding of the network's behaviour and the relations between actors/nodes in various problems (e.g. social networks) will be easier by grouping nodes into partitions. "*A partition of a network is a classification or clustering of the vertices in the network such that each vertex is assigned to exactly one class or cluster. Partitions, therefore, are very useful for making selections from a network to reduce its size and complexity*" (De Nooy et al., 2005).

Partitions are used to reduce networks in three different ways:

- Extract one part (local view). The simplest way to extract a sub network from a network is to select a subset of its vertices and all lines that are only incident with the selected vertices.
- Shrink each class of vertices into one new vertex (global view). To *shrink* a network, the only operation needed is to replace a subset of its vertices by one new vertex that is incident to all lines that were incident with the vertices of the subset in the original network.
- Select one part and shrink neighbouring classes to focus on the internal structure and overall positions of this class (contextual view). To obtain the contextual view, all classes are shrunk except the one under analysis,

Although cliques (i.e. maximal complete sub network containing three vertices or more, where the term maximal means that no other vertex can be added to the sub network without destroying its defining characteristic) and K-cores (i.e. maximal sub network in which each vertex has at least *k* connections within the sub network) are used to explain attributes of networks, this techniques are commonly employed to create clusters. (De Nooy et al., 2005)

2.5 Data Mining Overview

Data mining can be understood as the extraction of hidden predictive information from large databases. This technology that has inspired an explosion of interest from both academia and industry is generally used for companies to understand the data they store or predict future trends and behaviours allowing business to be proactive. One of the biggest advantages of data mining is the opportunity to clarify problems that traditionally needed excessive time to be solved. (Thearling 2007)

The data mining process is based on massive data collection, pre-processing of the information with multiprocessor computers and final pattern recognition and modeling of interrelationships using data mining algorithms. Due to the development and price reduction of the storing systems, nowadays all companies are able to store enormous amounts of data concerning their business. On the other hand, multiprocessors allow the fast analysis of big sized files in a way it was not possible some years ago. Finally, data mining algorithms process the resultant data with two basic scopes: first the prediction of trends and behaviours (that can be used e.g. in targeted marketing) and second the discovery of previously unknown patterns. (see Berson et al., 2000)

In essence, all the existing data mining techniques can be classified into two different groups:

- Classical techniques: used since the early eighties.
- Next Generation techniques: developed over the last twenty years.

Due to its high interest in this study, both techniques are analyzed deeper in the next sections.

2.5.1 Classical Techniques

Based on Berson's (2000), the most common classical techniques are statistics, neighbourhoods and clustering, all used since the early eighties.

Statistics cannot be considered as data mining in a strict sense. Statistical mathematics is used to understand the phenomena that have generated the data, manipulate these data, discover patterns and build predictive models sharing objectives with traditional data mining. The evolution of computers and storing systems has allowed companies to save enormous amounts of data, making possible the search of hidden patterns from new sources of information. Although statistics are always related to big amounts of data, they were not conceived to process such big sized files and this is the main difference between statistics and data mining. Statistics can be a helpful tool in the comprehension of a database through the use of representations like histograms that provide information about the statistic distribution of the data (probability distribution) or regression methods applied in prediction of trends. Despite all this, new data mining techniques are more efficient to process the big files that companies have to handle nowadays. (Fields 2005)

Nearest neighbour prediction and clustering are among the oldest techniques used in data mining (see Berson et al., 2000 and De Nooy et al., 2005). They are very similar in essence because both of them are based on the idea of grouping like records together. Nearest neighbour is generally used in prediction, whereas clustering can be applied in the detection of outliers (i.e. atypical or extreme values), summarization or exploration of data as well. The relation between clustering and segmentation has increased the importance of this technique. Clustering could be defined as the process of partitioning a set of data or objects in a set of meaningful sub-classes called clusters which will help to understand the natural grouping or structure in a data set. It is an unsupervised classification in the sense that there are no predefined classes or clusters at the beginning. Cluster is a collection of data objects that are similar to one another and thus they can be treated collectively as one group (the definition of the word similar depends on the specific problem). New techniques deriving from clustering like e.g. biclustering (i.e. a simultaneous partitioning of the set of samples and the set of their attributes) are nowadays a matter of study taken under the data mining perspective. (Busygin et al., 2008)

By applying clustering methods companies are able to e.g. group the population by demographic information into segments which they will use e.g. in targeted marketing.

Since one of the objectives of this thesis is to group cells into four groups regarding end-user's context, the use of clustering methods becomes relevant.

2.5.2 Next Generation Techniques

Developed over the last two decades, the new data mining techniques such as decision trees, neural networks and rule induction are gaining importance and their use is increasing these days.

Decision tree induction is a well-known discipline in Machine Learning. Based on the decision trees used in decision analysis theory, a decision tree or classification tree in data mining is a predictive model that can be viewed as a tree. Each branch represents a classification question and every leave of the tree are partitions of the dataset regarding those classification questions. One of the main advantages of this technique is that it divides the data on each branch without losing any information. Trees have been applied in business for segmentation or decision making, but they are commonly used for exploration (by simply checking the predictors and values utilized for each split of the tree), data pre-processing (for other prediction algorithms) or prediction (exploratory analysis). This versatility is the reason of the wide usage of decision trees in data mining. (Llora et al., 2001)

On the other hand, neural networks are a branch of the artificial intelligence focused on the development of networks that try to emulate the neural tissue of the brain. Although it is a relatively old technique (they were mentioned at first during the Second World War) they have significantly evolved since the middle eighties. The main characteristic of these networks is that they are capable of learning by themselves through experience (training with specific predefined patterns). There are several algorithms to model the neural networks (e.g. Kohonen or Back-Propagation networks). The choice of the specific model and the datasets provided in the training stage determine the future behaviour of the network. Neural networks have been used in business for a wide variety of applications (e.g. they have been applied to bank's applications that detect fraud in the usage of credit cards or decide whether a person asking for a loan is a potential debtor). In addition, they can be used for clustering and prototype creation, outlier analysis or features extraction as well (i.e. the detection of the most relevant predictors that will be used in building models), extremely useful applications in data mining. (see Krieger, 1996 and De Veaux 1997)

Machine learning can be considered as a subfield of the artificial intelligence (Ryszard et al., 1983 and Alpaydin 2003) because it usually refers to the changes in systems that perform tasks associated with artificial intelligence. Such tasks involve recognition, diagnosis, planning, robot control or prediction, tasks commonly linked to the artificial intelligence theory (Nillson 1996). The objective of machine learning is the design of algorithms that allow computers to be able to predict and learn through experience. This need comes from the difficulty in programming the algorithms to solve some problems that cannot be focused on a classic way. These problems include e.g. those in which there is no human expertise to know the decisions to take considering the obtained results, those where there is high human expertise that cannot be easily explained or taught (e.g. hand-writing recognition or natural language understanding), tasks and problems related to fast changing phenomena (e.g. finance and stock market) or those related to personal applications (e.g. spam filters that have to adapt to specific restrictions for every user). Machine learning is close related to the fields of statistics, data mining and psychology but with differences of emphasis. The objective of machine learning is the development of an accurate and effective algorithm embedded in the resulting computer system carried out to process the data, extract conclusions and take decisions over the results (Dietterich 1999). Pattern recognition, another important goal of neural networks and a subtopic of machine learning is an important aspect of data mining as well (Bishop 2006).

Finally rule induction can be considered as the most common form of knowledge discovery in unsupervised learning systems (see Berson et al., 2000). Besides this, it is the technique that fits better with the concept of data mining as the search for "gold" through a vast database. These rules are generally quite simple (i.e. then-if conditions) and they tend to be associated to relative values calculated from the data (e.g. percentages of usage, sales or repetition).

The business value of rule induction techniques lies on its simplicity. This technique can be highly automated and it is, therefore, a very easy to use method for companies. But the if-then rules present an important disadvantage: the overabundance of interesting patterns discovered make the prediction process even more complicated. Rule induction can be applied in concepts such as the return of the investment.

Although similar in the basic idea, decision trees and rule induction present important differences. The main distinction between them is that decision trees are stricter in the process of splitting the data when creating the branches in the way that every record matches one and only one condition. On the other hand, rules induction leaves the possibility of keeping data not classified in the dataset at the end of the process or have these data classified regarding different rules (there may be many rules that match a given record).

Rule induction technique is probably the most creative process used in data mining. It starts the classification from a high level rule and from that rule it begins to add constraints making the groups narrower and reducing the coverage. The clusters become more consistent as the rules are stricter creating strong groups from the data. Because of its simplicity and versatility, the rule induction is one of the most used techniques in data processing.

2.6 Network Visualization Tools

In the next sub-sections an overview of the evolution of network visualization techniques and tools is presented together with a comparative analysis of some of the tools available nowadays. The last chapter provides a brief explanation of the tool that fits best in the purposes of this thesis giving objective justification.

2.6.1 Overview of Tools

Due to the continuous increase of the volume of data that big companies store, the representation of the information using graphic tools has become a difficult task. Most of this information can be interpreted as networks where the nodes are objects and the edges or arcs represent the relations between these objects. The relations can be defined following many different purposes and approaches (e.g. when nodes represent individuals the connections can be social connections, likings or parenthood), they can represent physical measurements, computed aggregated (e.g. mean or variance) or abstract quantities (e.g.

probabilities). Besides, the connections can be directed, undirected, time varying or static. (Eiek 1996)

One of the most used techniques to visualize networks is the diagram of nodes and links. The graphical objects are placed in the display as well as all the connections among them. Thickness of lines, colours and sizes of nodes and links are used to provide additional information as encoded statistics associated with them.

The need of working with large networks involves the use of excessive information. The network visualization tools are overwhelmed with huge volume of data that they are not able to represent in a clear way, becoming cluttered and visually confusing. There are three reasons for this:

- Display clutter: for the reasons exposed, the tools are easily overwhelmed when processing too much data.
- Node positioning: The interpretation of the results depends strongly on the node positions.
- Perceptual tension: There is a tendency in the viewer of networks to consider closely positioned nodes as related. Conversely, distantly positioned nodes are perceived as unrelated. Nevertheless, lines connecting distant nodes are perceived better because they cover a bigger part of the screen. All the networks visualization should consider the perceptual tension to take advantage and use it to provide more clear images.

New tools solve the information overload problem through a clever node positioning, the use of colours, sizes, cluster methods (in order to reduce complexity and provide hidden information of relations and groups) and 3-D graphs, where the connections don't cross each other (see Eiek 1996).

2.6.2 Comparative Analysis

The table below presents a comparative analysis of some of the tools used for network visualization. They have been chosen from all the available tools in the market because of their features, which fit in the purposes of this thesis: a clear visualization of the context. In the table, the most relevant characteristics are shown after a previous analysis of every tool (all except the last two that are private and therefore not available for free). Aspects like "easy-to-use" or quality of the documentation although subjective, have been justified following objective criteria (e.g. how easy is to draw a simple graph without any previous knowledge is the argument to decide whether a tool is easy to use).

FEATURES **TOOLS**	**LANGUAGE USED TO DRAW A GRAPH**	**EASY-OF-USE (EASY TO DRAW A SIMPLE GRAPH)**	**APPLICATIONS**	**SUPPPORT**	**SPECIAL REQUIREMENTS**	**PRICE**
Jgraph	JAVA	No	Computer networks, SW. Architecture or database connections. CONCEPTUAL	Java API	JAVA (SDK)	FREE Although there is a PRO Version to purchase only
GraphViz	DOT	Yes	SW engineering, networking, databases, knowledge representation, and bio-informatics but NO large networks	DOT Language Manual	JAVA (SDK)	Free
Condor	Written in JAVA. MySQL commands and interaction with the GUI	Yes	Visual maps of social networks, Web site link structures, and concept maps of unstructured documents, online forums, phone archives, e-mail networks…	It is hidden the way it works (GUI)	JAVA (SDK) & MySQL	Free
Pajek	"Pajek´s"	Yes	Basically developed for Large Network analysis (Social Networks among many other)	One book explaining the concepts of social networks through this SW.	TXT TO PAJEK TOOL	Free for Academic purposes only

InFlow	Private Tool	Yes. As private Tool it provides an easy GUI, document-ation and company support.	Team Building, Organization Design, Internetwork Design , Post-Merger Integration, Knowledge Management , Leadership Development , Mapping Terrorist Networks , Industry Ecosystem Analysis , Network Vulnerability, Assessment , Community Economic Development, Discovering Communities of Practice , Mapping and Measuring information Flow...	Yes	None	Unknown the exact price. Three options of purchasing: 1)SW only with on-line Documentation 2)same plus introductory SW training 3)Consultant's Package for no experienced users in network analysis including a basic course
TouchGraph	Private Tool	Yes. Same case as InFlow	See the big picture of the data. Discover clusters and interrelations within data. Analyze and create reports. Discover patterns and refine the results.	Yes	Unknown	Unknown

Table 1 – Comparative analysis of network visualization tools (see Tools 2008)

JGraph requires big training in the tool before understanding how to draw a simple graph. On the other hand GraphViz uses a language (DOT) to write the files containing the network information that makes the process of drawing a graph a complex task. Condor's graphic user interface hides some operations needed in the process of context modeling. Finally, InFlow and TouchGraph presents a big number of advantages but they are private solutions.

Pajek tool deserves special attention since it is the chosen tool to present the graphic results regarding end-user's context due to its simplicity.

Pajek is a program for analysis and visualization of large networks (i.e. networks of more than thousands or even millions of vertices). It is a free available software (for non commercial use) running in Windows OS created on 1996. With *Pajek* it is possible to represent computer transportation, communication, social and intra - inter organizational networks, genealogies, flow graphs or molecule maps among many others. Due to this, it is a very versatile tool useful in many areas. The strength of *Pajek* resides in the way it works and processes the data. Large networks cannot be treated efficiently using standard network analysis tools which are mostly based on matrix representation (as e.g. *CondorView*) and are, therefore, limited to networks of reduced size (some tens or hundreds of vertices at most). This is the main reason to choose this network visualization tool in the representation of end-user's context from all the options. (De Nooy et al., 2005)

The goals of this tool are:

- Abstraction: by recursive factorization of a large network into several smaller networks that can be treated further using more sophisticated methods. This is based on the concepts of cluster and neighbourhoods. With *Pajek* it is possible to find clusters and extract and show vertices that belong to the same cluster separately.
- To implement a selection of efficient algorithms for large networks analysis.
- *Pajek* works with six different data structures to implement the algorithms:
  Network (a collection of nodes/vertices and links/arcs/edges), Permutation (reordering of vertices), Vector, Cluster (a subset of vertices grouped following some specific criteria), Partition (structures that contain information for each vertex such as which cluster it belongs to) and Hierarchy.

The usefulness of this tool will be shown by graphic examples in the last chapter.

3 Context Detection Algorithm

3.1 Introduction

This part goes further in the ideas and theories explained in the background chapter. From the study about context and location modeling done in this section, important insights will be derived that help in the design and implementation of the context detection algorithm.

As it was presented in the first chapter, one of the main objectives of this research is to build an algorithm for processing big amounts of handset-based data. The algorithm should be able to identify the end-user context. With this method academic usage-level studies can be carried out with contextual perspective in mind.

First of all, the concept of context is defined. For this purpose, all the ideas and context classifications analyzed on the previous chapter will be the starting point. The main aspects of context, known as context variables, will be described as well.

After that, the study goes further in the location modeling based on raw data. A big point of interest is how to use and analyze the data for this aim. In this section there is an explanation about the basic variables used, their relevance, and the possibilities that this type of analysis leaves open.

Finally, the visualization tool and the methods used for presenting graphic results are briefly described as a helpful value added that can show visually the results obtained in a numeric format.

3.2 Definition of Context

One possible way to find a definition of the concept of context is to ask the proper questions about users and context. This thesis studies mobile end-users and devices.

Considering end-users, the information required for detecting context and context variables is provided by the following questions:

- *Is the user moving?*
- *What date/time is it?*
- *Where is the user?*
- *What is the user doing?*

Concerning the devices,

- *What application is launched at this moment?*
- *Which base station is the mobile phone connected?*
- *What kind of connection is the mobile phone using?*

Based on some of the concepts and classifications explained in the second chapter, the questions mentioned above can be quickly translated into context variables as the following figure presents:

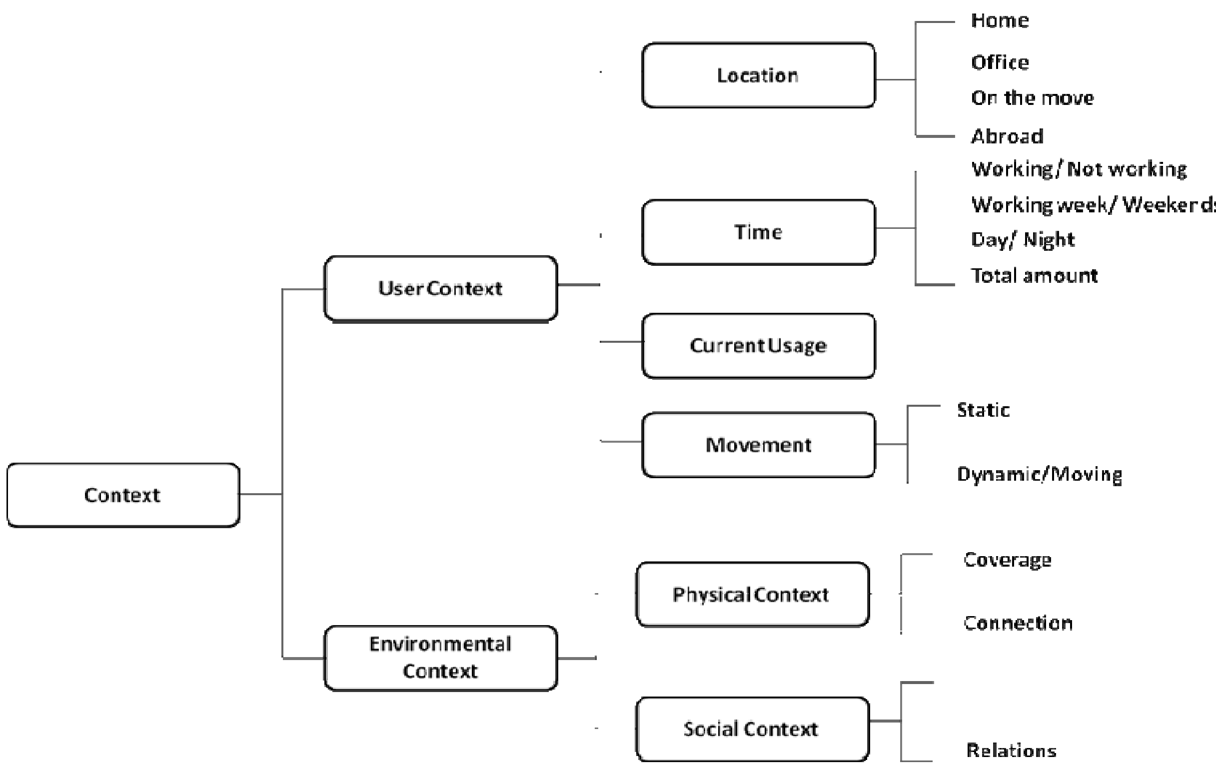


Figure 5 - Context classification

Before explaining the classification assumed in this study, a definition of context is now possible. For this study, **context** is "*a group of variables, parameters or characteristics that sets, specifies and helps to understand in a clear way, without ambiguity, the situation of an entity (1) in its current environment (2)*".

As a definition carried out for the purposes of this research, it will be explained part by part.

(1) Context is a group of variables parameters or characteristics that sets, specifies and helps to understand in a clear way without ambiguity the situation of an entity [...]

Thus, a group of variables, parameters or characteristics defines what is going to be used to detect the context and, later, the location of a mobile end-user (considering location as main part of the context).

As already mentioned above, context can be divided into two primary sub-contexts regarding the user (user or personal context) and the environmental context that relates to all the external factors where the user has no control (environmental or physical context in some studies e.g. Schmidt et al., 2005 or Chen et al., 2001).

For describing the end-user context, four variables have been identified:

- <u>Location</u>: that is not only considered under a geographical or positional point of view, but also as a specific way of using the mobile phone. The basic location will be marked as "*home*", "*office*", "*on the move*" or "*abroad*".
- <u>Time</u>: a very important aspect to take into account in context modeling. Time stamping allows the identification of context by means of e.g. dividing the day into blocks (night time and daytime or working hours and not working hours), dividing the week into weekends or working week, or considering the total amount of time spent (relative time indices in percentages).

- Current usage: i.e. application launched at a certain moment, access technology used or time of use among others (e.g. entertainment tool tells about the context in the same way agenda does – work or spare time).
- Movement: End-user context can be static or dynamic (e.g. "*home/office*" or "*on the move*").

The second part of the definition relates to the environmental aspect of the context,

(2) *... in its current environment.*

This sub-context can be divided into:

- Physical context: regarding to the handset, the physical context is described in terms of coverage (can be measured by detecting consecutive transitions between two specific base stations) and connection type (e.g. WiFi) terms.
- Social context: describes likings and preferences of e.g. music artists, applications used or websites visited and interactions between people (regarding voice calls or connections because of similar likings).

In this research, the data processing and visualization results are conceived under the hypothesis of the end-user context division into four basic contexts:

- Home
- Office
- On the move
- Abroad

Although this classification derives exclusively from the location perspective, the context "home" does not mean only a location, but also a way of using the handset in that specific area. Contextual information from the handset-based data is extracted taking into account several factors (cell identifiers, timestamps of the actions reported, applications launched, etc). This information is transformed into “specific” locations (in terms of base station identifiers) in the following step of the procedure.

In addition to the extraction of high-level contexts (the concept of cluster will be introduced later), also sub-contexts inside these high-level contexts can be found. Furthermore, the identification of usage patterns, trends and user habits becomes possible when using extracted contextual information with other data (e.g. application usage logs). Academically, the contextual study framework opens the door to many new opportunities (e.g. it is possible to analyze the adoption, usage and locations of use of new technologies across contexts).

Due to policy issues, privacy and end-user’s anonymity protection, it is not possible to “translate” these locations into geographical information with the data provided, so the detected contexts cannot be associated to a point in a map. The visualizations expected will show the movements of the mobile end-users (paths of base stations), the connections between base stations and the key contexts detected, but it will be not possible to extract geographical information from the base station identifiers. Nevertheless, there is always a chance to turn these base station connection graphics into geographical maps with the cells superimposed once the relation between base station and position is known.

Thus, base station indexes, the closest parameter in the dataset to “geographical” information available, are used as zone markers where the user spends his/her time. Obviously, a cell identifier is not the same as the assert “the user is at the coffee shop closest to his house in his/her neighbourhood”, but it is a first very good approach that linked to time stamping allows the researchers to identify basic places as the home/office/on the move/abroad mentioned above.

The process of modeling of the mobile end-user’s location cannot be fully understood without an explanation of the data used. These data will be analyzed in the following point.

3.3 Data Model

The method utilized in this thesis is a handset-based end-user research developed in collaboration of Helsinki University of Technology and Nokia. It consists of two main modules. The first one is a Symbian handset client that is basically running all the time as a background process, monitoring several phones' activities of interest, generating log files with this information and finally sending those reports to centralized Internet servers (usually one or two times per day). These specialized database servers are the second important module. The aggregated raw data derived from Internet servers are used for analysis and reporting.

Thus, after the users have signed an agreement contract, all the information regarding e.g. applications used, movement, voice calls or subjective information (direct answers to questionnaires) is stored.

There are two basic observation categories:

<u>Objective information</u>:

Applications and services usage as the following,

- Voice: voice calls log or phonebook contacts.
- Messaging: Standard and multimedia messages and Email (the number and form of attachments is also stored).
- Imaging: Photos and videos taken.
- Applications: activation, installation and removal information.
- Browsing: Websites browsed.
- Traffic: Package data traffic generated.

System information,

- Cell Id metrics.
- Battery levels.
- System crashes.
- Memory status.
- Profile actions
- On/Off report.

One important matter is that each one of these observations is stored with a time stamp, something that will be really useful as it will be explained afterwards.

Subjective information: Web based complementing questionnaires and on device prompted context sensitive "mini questions" are part of the initiative as well. Questionnaires regarding demographics and current context are used as well (mobile questionnaires).

Naturally, the size of the data logs will vary depending on the technical configuration of the research tool and on the device usage (typical sizes from ten kilobytes to one hundred and fifty kilobytes per day per user).

For researches, the advantages of this kind of information recovery and storage system are numerous. The big potential of this application is not only the fact that this kind of data overcome the problem of end-user's subjectivity (e.g. this method is more accurate questionnaires), but also the possibility of getting a new type of information (e.g. size of the files downloaded, precise music artists play lists or exact report of the base stations visited every day) that allows researchers to raise new questions and opens new ways of investigation or offers companies new business opportunities.

From all the data files and all the possible information available, this chapter is focused on the parameters that allow the context modeling.

First of all, the framework of this study should be mentioned. The data used for the modeling process refers to mobile end-users. When these users are moving, their handsets are changing their base stations all the time (i.e. access points to the networks). Typically, the mobile phone will be connected to the closest base station geographically or with the one with the best radio link quality available. When the user starts moving, the handset will check for another base station in order to receive the best signal. If the phone finds a better option, it will connect to this one and a change of base station will occur.

The information contained in the file is a time stamps sequence (date and time) that reports every change of base station along the day. Information about the new base station is provided as well. This file has a specific format with the information separated into columns.

The most remarkable fields are the following:

- Terminal id: Information regarding the end-user.
- Date and time: time stamp of every connection to a base station.
- Mcc and Mnc: Mobile country code and Mobile network code for the present situation (at the associated time stamp).
- Lac and Cid: Local area code and Cell identifier for the present situation. The information of these fields is hashed in order to maintain end-user's anonymity.

How to process the data to obtain context information is the main question that next parts try to answer.

3.4 Design of the Algorithm

The objective of this section is to describe a way to model end-user’s context that includes data processing and the design of an algorithm capable to detect context information.

From all the fields that compound the data analyzed, only terminal id (to identify the users), date and time (time stamps will be the key to detect context), cell id (to identify the cell where the user is) and mobile country code (to check user's movements abroad) variables are used.

The concepts of context and location are very close for the purposes of this study although location has been considered and defined as a part of the context. The steps to model the end-user context are, first, to try to identify big groups of cells and, second, to detect whether these groups can be classified as *home*, *office*, *on the move* or *abroad*.

3.4.1 Clusters

The big amount of data to process involves the grouping of cells into clusters in order to reduce complexity. But the grouping of "physically close" cells cannot be only considered from the context detection point of view. The visualization of the context through network visualization tools requires it to simplify the number of base stations that one single user visits along the time of the study.

How to identify clusters? First of all, a cluster is going to be considered as a group of cells that are close physically. There are situations where the boundary of separation between two cells is in the middle of a building (e.g. a home) or others, maybe more realistic, where there is a home context in the overlapping zone of coverage between two base stations. And this is a more realistic situation because the coverage of close base stations generally overlaps in their limits. The reason is evident: the zones without network coverage have to be minimized in order to provide service at every place. If the house is in the mentioned overlapping zone, every time the mobile user goes e.g. from one room to another a change of base station occurs because he/she is crossing this boundary. Another transition will take place every time that there is interference or a power loose among other situations. Of course, the context is the same in this particular example (home) and the context detector algorithm should be able to understand those situations grouping both cells as one.

With the data available, it is easy to identify these boundaries. In the log files there are many situations where it is possible to see a sequence of changes between two particular cell ids, changes that are very short most of the time. In the report file, it can be observed that e.g. from one base station marked with the cell id “A” there is a transition to a base station marked with the cell id “B” and subsequently another change to the first one (“A” again). These transitions usually appear many times spread in the file, what means that they are not just a bug because of a dead battery (e.g. a possible bug could be a situation where the mobile phone connects to one base station, the battery dies and when the mobile phone is switched on again it is done by chance in the first base station) but they are a symbol of cells that can be grouped into one.

The cluster identifying is the first step in the context detection. The algorithm developed works with group of cells. This assumption has two big advantages: first the data reduction and second the conceptual aspect.

The data reduction (grouping all the cells that follow the rule explained above) results in a better visualization (although this aspect will be explained after, less cells means less nodes and less connections what clears the image) and in a more efficient processing (the algorithm to process the data will be faster when handling lower amounts of data).

Clusters have to be understood as areas that can include several cells. This is one of the reasons why it would not be possible to translate exactly contexts into locations (the other is the impossibility to know real geographical positions of the base stations, coded by the network operators). *Home* context is going to be identified with the cells that covers real home and those which cover its surroundings as well if e.g. every day the mobile user is going to visit and have a brief talk to his/her neighbour or if this user simply goes to fill the tank of his/her car in the gas station closest to his/her home. Clustering is especially important in cases like office context where people move more often along the day.

At the end, all the cells are grouped into clusters. These clusters will consist of a non determinate number of cells or, in some cases (generally for most of cells detected as on the move context), the cluster will be formed by a single cell.

3.4.2 Contexts and Sub-Contexts

As already said, home and office cannot be understood as geographical positions exclusively. In this study, home represents not only the place where people reside but also specific locations where people spend long periods of time at particular hours of the day. For this reason, from the very beginning the idea of context was broken down into context and sub-context.

With the sub-context concept (i.e. a more detailed part of the primary context) it is possible to find situations where there is more than one home or office for the same end-user (sometimes even more than two or three). The explanation is simple: e.g. for one person his/her real home can be identified but if this person spends every weekend at his/her parents place, this location will be identified as home as well. Therefore, the sub-context is a division of the main context (home or office) into several little contexts that fulfil the rules used to identify the big context but that cannot be considered the same location. The context detection algorithm ranks all the different homes and offices detected taking into account the total time spent at that position.

The need for sub-context appears in the research objectives of this study: it is more important to place and identify where the mobile user does what he does in terms of context (e.g. use a specific application at home), rather than translate these places into geographical positions. Of course, the idea of personal and focused advertising is very attractive for companies for obvious reasons. But from the researcher and company point of view, user habits knowledge is a more powerful tool that can be used in studies related to adoption of new technologies (regarding penetration) or in a better segmentation of the market.

The case “abroad” is special and it shows the real essence of this understanding of context, so close to the term location. For the interest of a service usage analysis, it is less interesting to follow the user movements when he/she is abroad than being able to identify the moments when this user is in another country (and all the information regarding his/her interactions with the smartphone). Under a location perspective, the context abroad is reduced to one single cell for every country visited. The idea of showing the path to the airport or analyzing the usage abroad prevails over identifying the moves when the user is travelling. The big context here is “abroad” and the sub-context, the name of the specific country visited.

The limitations of this study begin at the point that subjectivity is needed and becomes an important part of the decision making. But, although subjective, the rules to detect the context have a simple and certain logic in this particular case at least (home, office and on the move contexts) as it is explained later.

3.4.3 Time-Based Modeling of Contexts

In the background chapter there is a list of previous studies where time is considered as an important contextual factor in general and a very important parameter to identify contexts (see Chen & Kotz, 2000 or Gross & Specht, 2001). From all the variables used to define the context, time is the key of this research in the detection of end end-user contexts.

The idea of using time for modeling the context is just a practical issue: every action in the log files used has a timestamp that allows the time analysis. Although there are no more chances left than using time to process the data with the information available, time would be always the best solution for the process of context modeling. The aim of the algorithm is to identify the context of every cluster from among *home*, *office*, *on the move* or *abroad*. For this purpose, time is used to classify user behaviour into several statuses, regarding three different approaches:

- Division of the week into working week and weekends.
- Division of the day into day and night.
- Division of the day into working hours and rest of time (duty and off duty).

The total amount of time spent in every one of these situations is the value used in the analysis.

People spend most of the time at home. This idea can be translated into terms of absolute times. To detect home context, a time-based modeling identifies those clusters where the mobile users usually sleep, those where they spend most of weekend hours or those with the biggest amount of time spend during non working hours. The limitations here can be found in the subjectivity that appears when deciding the threshold values (e.g. when the working hours start and when they finish) but not in the assumptions made. It is true that there are many exceptions and that every user has a very particular schedule. Therefore, the difficulty here is to design a general algorithm that has to be powerful at the same time.

From this conception of time as a key parameter in context classification, it is now possible to understand the definition of context presented before. To use the variable time to divide the user behaviour into blocks provides information about leisure time (home), working habits (office that could have been named as university as well) or actions and activities done in temporary locations (on the move).

Finally, it is possible to assert that the value added of this method regarding segmentation is very high. Companies are more interested in knowing e.g. what kind of applications end-users use more often at every one of these contexts: while working, at leisure time and on the move (in the way to somewhere).

3.5 Context Detection

All the necessary elements to detect context have been presented along this chapter. First the clustering method that makes possible the grouping of close cells (in terms of context) and second the use of time as the key variable in the identification of the type of context.

The algorithm developed to process the data and detect end-users context is explained in this section. This algorithm has been implemented with Java but the interest is focused in how it works (high level description or pseudocode), the rules to detect the clusters and, after that, the conditions to identify the context. In other words, the chapter introduces the design of the algorithm.

First of all, it is necessary to clarify some aspects related to the dataset. The file that contains the data used in this study is a very big sized file. It stores the data regarding all the end-users what makes this file something difficult to process for a normal computer. Because of this, the first practical step is to split this big file into as many files as users it has. This action lets the whole process goes faster.

From now on, only one user will be considered at the same time. Therefore, it is correct to say that the algorithm processes every user individually.

The file (that is sorted by users, date and time) will be modified during the whole process, adding and erasing columns with all the calculations done. The file processing is the cornerstone of this research. For this reason, special attention is paid to these changes in the format of the file.

The first thing to do with the file is to select those users with enough data available (for example end-users with less than three weeks of active usage will be excluded). The reason is to have enough data to detect contexts and not to give the same importance to contexts detected for users with very few information.

Once the users that will be finally analyzed are detected, the split of the file begins. In this process a new column is added: the day of the week. This information will be used later in several calculations and functions (it can be considered a part of the timestamp).

One of the first actions is the **detection of the context abroad**. With the specific column containing the mobile country codes, it is possible to extract all the situations when the end-user moves to other countries. The main drawback at this point is the fact that the dataset can have mistakes that should be detected e.g. cases where a different mobile country code appears for just a brief period of time (seconds). All these cases must be erased as well as every movement abroad with less than ten actions (at least ten timestamps in the country visited and more than six hours spent on it will be required to consider that the user under analysis is effectively out of the country). Once the context is detected, all the cells of the same trip are grouped. This action will have particular effects that are described in the following sections.

The data provided can have duplicated cases not only because of the abroad context detection (the group of cells have the same cell identifiers). In this study, a duplicated case is a sequence of repeated cell identifiers for consecutive timestamps. These situations cannot be considered bugs, but they must be erased. The explanation of these repetitions is simple: when the system of the base station falls down, the phone suffers a power loose or simply when this phone is switched off and switched on in the same position, the cell identifier of the following case is the same. To detect context the only thing needed is a sequence of different cell identifiers. This sequence is “the path” that the user follows every day and that has been registered in the log file. With this sequence, time spent at every location is calculated to detect clusters, and due to no more information is required, those repetitions are irrelevant. It is possible to erase them without losing any information just preserving the first appearance of the repeated cell identifier.

3.5.1 Cluster Analysis

Next step in the procedure is the **cluster extraction**. Before this, it is needed to calculate the time spent in every transition and, with these times, the total amount of time spent for every user in the whole panel. Transitions (consecutive reports) of more than one day are not considered here. Thus, the maximum amount of time spent in one transition is twenty four hours. The reason is obvious and it is explained under the cluster point of view in the specific sub-section: long periods of time in a single cell could not represent “presence” of the user in that cell. They could be explained as e.g. “power loose” or “dead battery”, giving big importance to cells that are just “on the move”.

Once the absolute times have been calculated, the algorithm tries to find sequences “A-B-A” i.e. when from cell “A” the user moves to cell “B” and from there he/she goes back to “A”. In the file, this repetition will appear when three consecutive cases with this sequence of cell identifiers are found (three at least but it could be “A-B-A-B” and so on). This kind of situations that could be called “sandwiches” presents very valuable information. A repeated sequence of cells like that permits the extraction of user behaviour and user contextual information. It is possible to find a big number of examples that can explain these movements (e.g. a big house that is covered by a couple of base stations so when the user moves along the house, the mobile phone is connecting to one and changing to the other all the time; a person that works in the office and is moving all the time to his/her colleague room which is covered by a different base station; a mobile phone user that has his/her parents house in front of his/her own house and he/she makes visits every day) but in all of these cases the conclusion is the same: the context does not change even if the cell identifier is changing what does not strictly mean that the user is moving. These sequences of cells are the basic idea of the clustering method developed. But some kind of recurrence is needed before asserting that two cells can be considered close cells because the sequence could be a matter of chance. For this reason one time situations should not be detected. Another advantage is that due to this assumption the algorithm becomes stronger: the fact that two cells are together physically is not the only requirement to consider them as the same context. The number of times that a sequence of repeated cells has to appear to consider the cells involved as a cluster is a matter of analysis that will be discussed in the next chapter.

In the process of cluster detection, special attention is paid to the situation of the cells marked as “abroad”. Due to all the steps explained before, every travel to other country is reduced to one report (one single timestamp that can be otherwise very big regarding time spent on it) what means one case in the file. Although the chance of clustering a cell identified as “abroad” decreases as the number of repetitions needed increases, the clustering method developed ignores the clustering of these cells. It is not possible to accept the risk of grouping timestamps regarding trips abroad.

Cell id.	Total time spent	Cluster id.
A	100	A
B ←	5	B ←
A	100	A
B	5	B
C	50	C
D	10	D
A	100	A
E	15	E
B	5	B
D	10	D
F	30	F
G	20	G

Figure 6 – Input data model

The results of this clustering process are given in a new column added to the file. This column is initialized with the same values that the column cell identifier has. Every time a cluster is detected, the new column changes. The cells involved are grouped simply by changing their cell identifiers (every time their cell id. appears in the new column) for the cell identifier with the biggest amount of time spent. The result is e.g. that there will be no more apparitions of the cell named “B” in the new column (it will be replaced for “A”) if “B” has less amount of time spent than “A” (in the other case “A” is replaced for “B”). The detection goes from the first case to the last one. Thus, e.g. when every apparition of cell identifier “B” has been replaced in the file with “A” once they are considered a cluster, new situations “A-X-A” or “X-A-X” appears due to that change. For this reason a cluster usually contains a group of more than two cells.

Cell id.	Total time spent	Cluster id.
A	100	A
B ←	100	A ←
A	100	A
B	100	A
C	50	C
D	10	D
B	100	A
E	15	E
B	100	A
D	10	D
F	30	F
G	20	G

Figure 7 - Process of clustering (cell B has been grouped into A)

At the end of this process the new column created contains the cluster identifier associated to the cell that appears in the column with the information about the current cell. If the cluster identifier is different to the cell identifier, this cell has been grouped, at least, with the cell which has the same identifier as this cell's cluster identifier. If the cluster identifier is the same as the cell identifier, this cell has not been grouped or this cell is the most visited from all the cells grouped (this is why the cluster has been marked with its cell identifier).

It is important to highlight the fact that the apparition of a first cluster affects the rest of the clusters found because when two cells are considered close cells and are grouped, one of these two cells is renamed with the identifier of the other. If every apparition of the renamed cell is changed, this means that new clusters between the grouped cells and new cells can appear. This is one of the biggest limitations: due to a time-based modeling, the order of the cases is essential in the detection and determines the quality of the results.

After all the close cells have been grouped, the next step of the algorithm is to calculate all the times needed in the context detection,

- Total time spent during weekends.
- Total time spent during the working week
- Total time spent at nights.
- Total time spent at working hours.

These times are calculated for every cell and then for every cluster.

3.5.2 Context Extraction

With this information the algorithm starts the **context detection**. The aim of this process is to classify every cell used for one user into "home", "office" or "on the move" context, since "abroad" has been already detected.

The first action is to classify every cluster regarding to static or moving (home/office or on the move). The algorithm calculates the total amount of time spent in a cluster and establish a threshold of minimum presence on it. For all "on the move" clusters, the time spent will be lower than this threshold.

The procedure of the algorithm is as it follows:

1) To classify clusters into "home/office" and "on the move" (detecting big contexts) by checking the total amount of time spent on them.
2) To detect whether the resultant clusters are office or not (in that case they will be considered home) checking all the times calculated and establishing thresholds.

As already said, the second step uses basic but logic assumptions considering time to mark clusters of cells as home or office. Generally, a typical user spends at work eight hours per day on average, never sleeps at the office and usually does not go to work on weekends.

Although these requirements are very general, they have two big advantages: first they allow the classification of clusters and cells in the way needed and second they are very flexible (the limitations and the weakness of the context detection lies on the thresholds and the range of hours taken into account, something easy to modify). In the next chapter all these practical issues (values of the thresholds) will be analyzed deeper.

The method used here is sequential so it processes every line of the file that represents a specific moment of time and a position, extracts the cluster related to that position and finally determines the context associated to this cluster which will be not analyzed again (obviously there will be more cases with this cluster involved, overall if the cluster is an important one, but it only has to be processed once).

The first important action is to determine whether the cluster is home/office or on the move. For this purpose, the algorithm check the relative time spent on that cluster as a percentage of the total amount of time spent for this end-user in all clusters.

The algorithm marks the cluster as on the move if,

$$\frac{\textit{total amount of time spent on the cluster under analysis}}{\textit{total amount of time spent on the whole panel}} < \textit{threshold}$$

If the cluster under analysis does not fulfil this prerequisite and the mentioned index is equal or higher than the threshold established after, it will be considered as home or office. At this point, the algorithm follows checking whether this cluster is office. If not, it will be considered as home.

To mark the cluster as office, the code must fulfil all the following conditions:

$$\frac{\textit{total amount of time spent on weekends on the cluster under analysis}}{\textit{total amount of time spent on the cluster under analysis}} < \textit{threshold 2} \qquad (1)$$

This is a logic rule considering that a standard end-user does not usually go to work during weekends. Instead of giving a 0%, the algorithm tries to be "flexible" and consider possible failures of the data in that threshold (that should be low) e.g. when a battery dies. In this situation the last report of the handset regards one specific cluster even when the user is still moving. When the user turns on the mobile phone again it reports from the new cluster producing two kinds of failures:

- A connection between two clusters, not close physically, is created. This mistake could be solved by checking the positions more accurately (taking into account previous cases and using statistics).

- The cluster where the handset gives the last report is considered as the cluster where the user is until the next report. This means that the algorithm consider this amount of time spent between both reports as time spent on this cluster, what is false. The problem here is that an accumulated number of repeated mistakes on the same cluster could produce "fake" big contexts and confuse a very clear "on the move" context with "home" or "office". The big advantage of considering users with more than three weeks of data is that all these failures are minimized (the calculated times are bigger with when more data is analyzed so all these situations are considerably reduced).

$$\frac{\textit{total time spent during working hours on the cluster under analysis}}{\textit{total amount of time spent on the cluster under analysis}} \geq \textit{threshold } 3 \qquad (2)$$

In the second condition the working hours are checked. To consider the cluster under inspection as office the user should be on this location during working hours most of the time. The definition of working hour is subjective again and it is part of the next chapter to decide the thresholds in order to make the algorithm as general as possible without losing strength. Of course, and even being flexible (the algorithm can consider a range of more than ten hours as the working hours) it is impossible to take into account every particular case (e.g. night workers). This is part of the weaknesses and limitations of the detection method.

$$\frac{\textit{total time spent during night hours on the cluster under analysis}}{\textit{total amount of time spent on the cluster under analysis}} < \textit{threshold } 4 \qquad (3)$$

The last of the conditions to mark a cluster as office regards night hours. To establish this threshold, a standard user working with a “normal” shift is considered. The subjectivity of the threshold is again the weak spot of the code at this point.

If a cluster fulfils these three conditions all together, then the algorithm marks it as office context. If one of these conditions is not satisfied, the cluster is marked as home.

The algorithm creates the sub-contexts simply by enumerating the contexts found. Every different home and office context detected is numbered (sub-context detection) and sorted at the end considering the most important home (home-1) as the home sub-context with biggest amount of time spent on it. For the clusters marked as abroad, is in this part of the code where the algorithm creates the context ("Abroad" and the number of the country visited, sorted considering total amount of time spent) and the sub-context (as the name of the country visited).

As it was explained before, it is important to separate sub-contexts in order to identify all the different environmental situations of a specific end-user. In a normal situation, a user has more than one place where he/she is not working at all and all the applications related to infotainment take more relevance. All these places could be considered as another households (home sub-contexts), parts of the big home context.

The situation in the case of the office is very similar because it is possible to find people working and studying and both sub-contexts could be marked as office.

The last step of the algorithm is to prepare the data and create the network files used for the visualization software. In this point, the algorithm prepares two new samples of data from the original one to be analyzed as well: data of working days and data of weekends. The importance of separating the data following this procedure is very clear: a typical working user has different behaviours during the week and at weekends. Besides, it will be possible to test the strength of the detection considering e.g. that a typical user does not usually go to work during weekends.

3.6 Implementation of the Algorithm

The programming language chosen to implement the algorithm is Java. Despite the numerous reasons of using Java instead of any other language involving execution times, efficiency, resources utilized or the advantages of working with the Java Virtual Machine, the interest in using Java is the big Application Programming Interface related to file processing. Basically, the method implemented has to process the file and e.g. read it or write it adding new columns. Finally this method processes the information inside doing basic calculations.

Although there are many other languages specifically developed to process text files (e.g. PERL), the Java code is more clear and understandable what makes it a suitable language to implement a code that is supposed to be used in later studies related to e.g. segmentation or new technologies adoption.

3.7 Visualization

One of the objectives of this thesis is to visualize the extracted information. It is unavoidable to link the concepts of context visualization with the geographical position but, as already said, it is not possible to translate the coded base stations into geographical locations. Nevertheless these graphs can provide very important and valuable information related to context.

The data files have been already discussed. They consist of a big number of rows and columns regarding end-user’s movements. Even if it is not possible to extract geographical information from them, these visualizations will show all the “paths” that a user follows every day. The purpose of the visualizations is to help in the context detection by giving a new type of information. The advantage of a graph is that it offers the possibility to understand a big mass of data at first sight using nodes, arrows, sizes and colours. The nodes are the cells, the arrows represent the connection between cells (movement of the user), the size of the nodes symbolize the importance of the clusters in terms of time (total amount of time spent for the user in a cluster) and the code of colours for the nodes shows the part of the day when this cell is mostly visited.

Pajek was presented in the second chapter as a suitable tool for large networks visualizations (nets of more than one hundred nodes). In the process of visualization, every cluster is considered a node in the network to model. With this large amount of data to work with, it is now possible to understand the need of the data reduction and the relevance of the clustering method.

Although the graphic results are not a geographical map, it provides information about movements (i.e. paths), relevant context environments (groups of connected clusters) or the most visited locations (represented with the biggest nodes). This kind of pictures are used to check if the context detection has sense (e.g. the environment of a cell classified as office should be, generally, green and orange because these are the colours assigned to the block of the day when people usually work at office).

The way this pictures help in the context detection can be easily shown in the figure below,

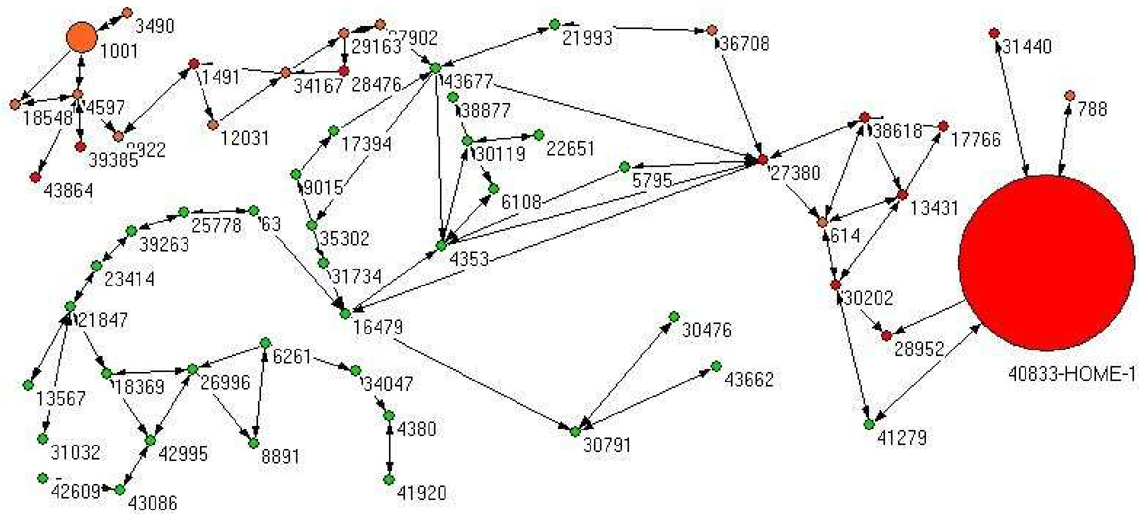


Figure 8 – Context visualization for weekends using Pajek

It is not a coincidence the fact that the biggest cluster is the one marked as “Home-1”. This example regards weekend days, so it is not strange either that there is no office context in the visualization (this is one of the advantages of separating the data into weekends and working weeks). Due to the procedure followed, it is not possible to find isolated clusters, what would have no sense (all the clusters should be connected if there is no missing data).

In the following examples situations where a user travels abroad is presented,

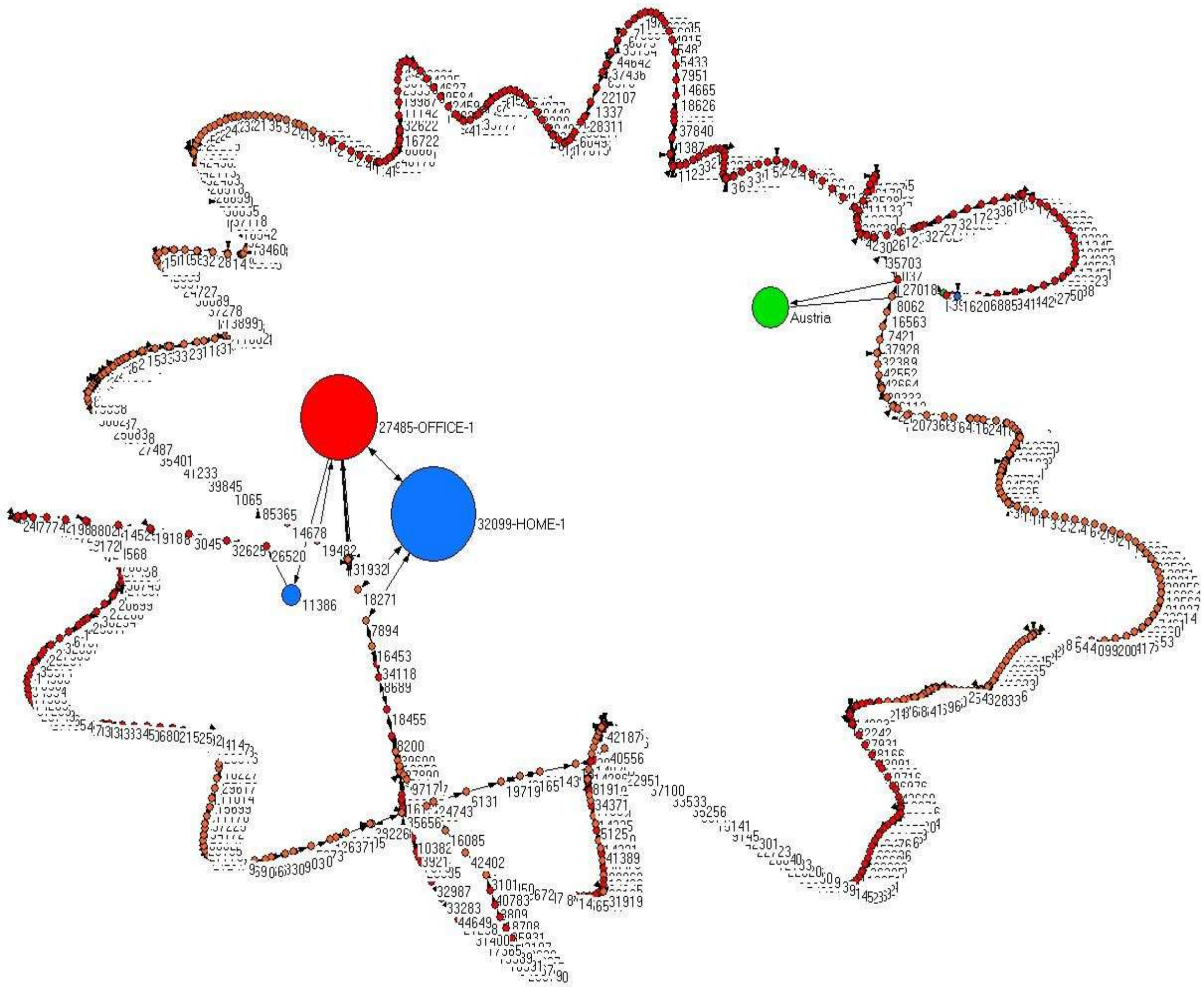


Figure 9 – Context visualization using Pajek – Example 1

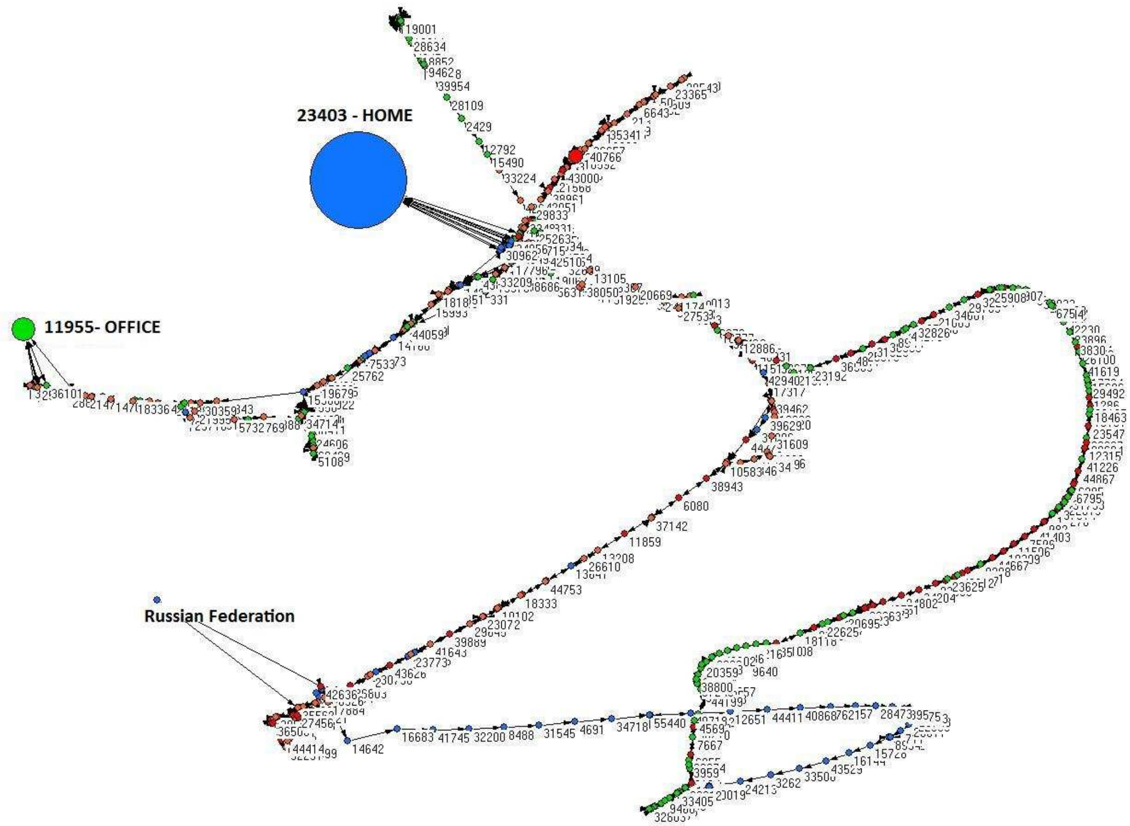


Figure 10 - Context visualization using Pajek – Example 2

In the figures above all the paths followed by the user in its everyday life can be easily seen as well as all the possible contexts detected by the algorithm: first home and office marked with these key words in the nodes, second abroad marked with the name of the country visited (in these cases are Austria and the Russian Federations respectively) and finally on the move (the rest of the clusters).

The sizes of the nodes reflect how important are the cells regarding to time spent for the end-user in them as already said. This is the reason why home, office and abroad (the green node identified as Austria) are the biggest nodes in the second example again. Important conclusions can be extracted from these graphs. For example, it is known that “on the move” clusters are smaller than the other because of the implementation of the algorithm. In the example presented it is possible to find at least a couple of “on the move” nodes bigger than the rest. They are probably sub-contexts like home or office that have not been detected because of the low accuracy of the algorithm.

On the other hand, the colours provide important information as well. The colour of the nodes informs about the most frequent hours of the day in which the user spends his time in the cluster represented by that node. The day is divided into four groups of hours:

- From 0-6 am (Blue)
- From 6-12 am (Green)
- From 12-18 pm (Orange)
- From 18-24 pm (Red)

After the analysis of times for every cluster (the algorithm calculates in which of this groups of hours the user spends more time at this location), every one of these nodes is coloured with one of the mentioned colours. The colours can be helpful to test the strength of the algorithm in the office context detection especially. Typically, all the nodes close to a cluster marked as office should be coloured as green or orange considering that a "standard" user stays at work in the range of hours from six in the morning to six in the afternoon.

The visualizations provide interesting information involving the context abroad as well. It is possible to extract paths that could be called e.g. "path to the airport" or "path to the port". This information could be easily deleted from both file and visualizations if only information about home or office is requested.

Thanks to this graphs, all the basic questions to test whether the algorithm works properly and detects contexts in a logic way can be asked and easily answered.

4 Application

This chapter is focused on a particular application of the algorithm developed. Using a specific dataset provided by Nokia, the study goes deeper in the analysis of the algorithm's procedure, the results, the critical variables of the context modeling and some other relevant aspects (e.g. visualization).

The objective of this chapter is to describe a new approach of context detection based on all the concepts examined in the background and test the efficiency and strength of the algorithm through the results obtained.

4.1 Dataset

The dataset used in this research has been presented along the chapters two and three. In resume, it is basically a text file containing a big number of rows and columns storing information about user movements (time stamps, location and user identifiers). All the data utilized comes from a Finnish panel. It can be considered a representative sample of the Finnish market composed of 643 users.

From all the fields in the file, the final input of the algorithm only requires the following columns:

- End-user identifier
- Date of the report
- Time of the report
- Cell identifier
- Mobile Country Code

The rest of the columns must be erased, as well as all the cases with a missing value.

The file has to be sorted by users and then in a chronological order (using the variables date and time). This is another important requisite of the input data thus the algorithm processes the file sequentially (i.e. line by line) and the context detection requires the manipulation of the timestamps what is easier if the information is arranged this way.

4.2 Configuration

All the variables used by the algorithm in the process of modeling are explained in this section. First of all, the thresholds of the rules to exclude users, detect the context abroad and identify clusters are introduced. The time-based assumptions considered to detect the rest of the contexts are described in the following sub section. Finally, all the parameters used in the process of context detection are analyzed.

4.2.1 Thresholds

In the previous chapter the context and location algorithm detector has been described from a theoretical point of view. This section is focused on the practical aspects explaining things like how it works when applying specific data or the definitive thresholds established.

The first step is the deletion of information about users with less than three weeks of data and every case with a missing value. From the original file providing data of 643 users, after the erase action there are 576 valid users left. This is a considerable number of users that permits the generalization of the results.

At this precise moment the algorithm identifies “abroad” contexts by checking the mobile country code. First of all, it analyzes the most common mobile country code. When the most dominant country is detected, every country code different to this one is considered as abroad. At this point, the code analyzes the first apparition of a different mobile country code, it checks first that the code is valid (if it exists in the list of valid mobile country codes), second that there are at least ten reports in that country and finally that the user has spent at least six hours abroad. Then, the algorithm changes the name of the cells involved over the trip.

Once the correct end-users have been selected, the algorithm proceeds with the exclusion of repeating base stations and the calculation of times spent in every cell (parameters needed for the clustering process).

At this point, the clustering process begins. As it was explained in the chapter three, clustering is a way to group physically close cells. The rule to identify close cells is to find repeated and consecutive changes between two specific locations in the file. The number of repetitions needed to consider them close affects seriously to the results and the number of contexts detected in the following steps. This will be described in the section 4.3.

4.2.2 Time-Based Assumptions

When the clustering process is finished, the algorithm calculates the grouped times needed in the context detection (the algorithm calculates them for the clusters and not for the individual cells). These times are, for every cluster:

- Total time spent.
- Total time spent during working week (from Monday to Friday).
- Total time spent during weekends (Saturday and Sunday).
- Total time spent during working hours (defining working hours as the hours in the range of time from Monday to Friday, from eight o'clock in the morning to six o'clock in the afternoon).
- Total time spent at nights (defining night hours as the hours in the range of time from Monday to Friday, from midnight to six o'clock in the morning).

It is important to remember the fact that the time-based rules have been thought to detect two particular contexts: home and office. The context detection proceeds as it was explained in the previous chapter: first it decides whether the cluster under examination is a static context (home or office) or not (on the move), and second it finds out whether this cluster is office or home. The rules to make the last step are based on basic assumptions regarding time:

- A standard user does not sleep at the office (unless he/she works at home)
- A standard user goes to work in the range of hours in between eight o'clock in the morning and six o'clock in the afternoon.
- A standard user does not go to work during weekends.

The weakness of the algorithm is not only in the dataset but also in the subjectivity and flexibility of the assumptions mentioned above (e.g. it is possible to find situations where the user works at home, goes to work during weekends or has a night shift).

4.2.3 Context Detection Parameters

The algorithm starts the context detection based on calculations done with the times explained in the previous section.

The rules for context detection were explained in the first chapter with theoretical thresholds. In this section, specific thresholds are provided. The first assumption is explained and in following sections, variations over these thresholds and its influence in the results will be analyzed.

The first decision is to mark the cluster as "home/office" or "on the move". This is done based on the following rule,

$$\frac{\textit{total amount of time spent on the cluster under analysis}}{\textit{total amount of time spent on the whole panel}} < 0.04$$

A four percent of the total time of one user means less than one hour of presence per day. So, to consider a cluster as “home” or “office” instead of “on the move” the user has to spend on that cluster more than one hour per day.

The clusters that pass the first filter are considered as “home” or “office” as yet known. The following rules determine whether a cluster is “office” or not (“home” in that case).

To mark the cluster as “office”, the code must fulfil all the following conditions:

$$\frac{total\ amount\ of\ time\ spent\ on\ weekends\ on\ the\ cluster\ under\ analysis}{total\ amount\ of\ time\ spent\ on\ the\ cluster\ under\ analysis} < 0.1 \qquad (1)$$

If the user stays more than the 10% of the time on this cluster during weekends, it is marked as “home” context. If not, the process continues checking the next condition. As suggested in the previous chapter, this parameter should be low considering that a standard user rests during weekends. However, the lower the threshold the less flexible code as it was explained in chapter three. The limitations of the algorithm are not only related to not considered user activities (e.g. weekend workers) but also to the possible data failures commented in that chapter.

$$\frac{total\ time\ spent\ during\ working\ hours\ on\ the\ cluster\ under\ analysis}{total\ amount\ of\ time\ spent\ on\ the\ cluster\ under\ analysis} \geq 0.75 \qquad (2)$$

Regarding working hours, the algorithm considers the cluster under inspection as “office” if the user spends on it the 75% of the total time during working hours at least.

$$\frac{total\ time\ spent\ during\ night\ hours\ on\ the\ cluster\ under\ analysis}{total\ amount\ of\ time\ spent\ on\ the\ cluster\ under\ analysis} < 0.05 \qquad (3)$$

Only a 5% of the total time spent on the cluster under analysis can be in the range of hours considered as night hours (from Monday to Friday, from midnight to six o’clock in the morning), what means 1.2 hours per day.

4.2.4 Sensitivity Analysis

This section is focused on the evaluation of the parameters used by the algorithm. One of the main points is to understand how variations over the values of the thresholds can affect to the number of contexts and sub-contexts identified.

It is important to remember some aspects of the dataset used in this study before extracting any conclusion. First, the file mentioned stores information of 644 Finnish users. Second, only users with more than three weeks of valid data are analyzed (this leaves 578 users from the original 644 users). Third, no missing variables or wrong values are allowed so these mistakes must be deleted in the pre-processing stage.

It the previous section all the conditions to identify and extract the user's context have been explained in detail. Basically, it is possible to summarize all the process in a decision tree as in the following figure,

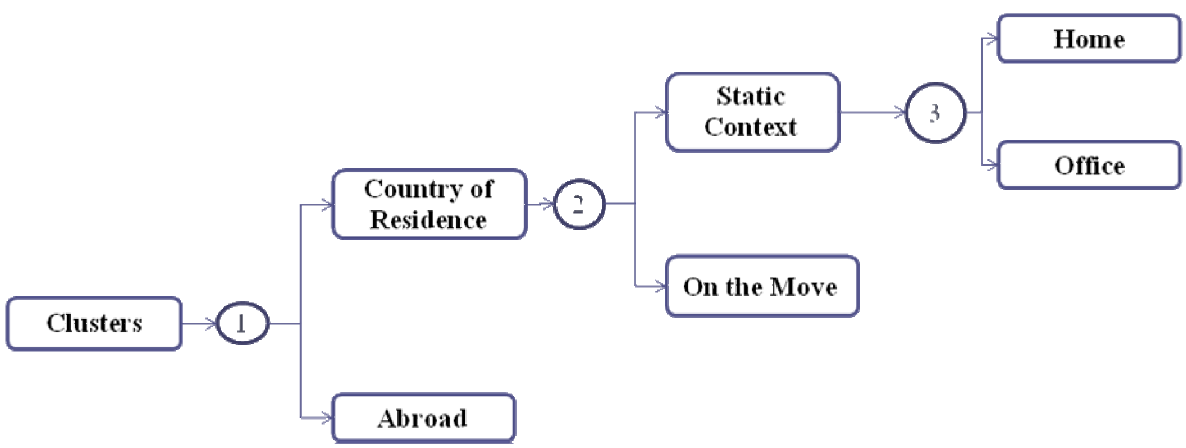


Figure 11 - Decision tree for the context detection process

Where every one of the decisions can be understood as one of the formulas explained in the sub section 4.2.3.

For the sensitivity analysis 21 initial trials testing different values for the thresholds have been run (the specific values and results are explained thoroughly in the Appendix B). The main findings after modifying every parameter keeping the rest static are:

- A raise in the number of “sandwiches” needed to detect clusters increases the number of clusters found, as well as the number of sub-context identified (a cluster and a sub-context are the same here). This action is essential if the objective is to detect the more contexts the better.
- The condition regarding abroad context detection becomes stricter if the algorithm checks the amount of time spent abroad together with the number of consecutive timestamps in a different country from the home one instead of checking only the last of this two requirements (e.g. the algorithm detects 112 travels abroad if only the timestamps condition is considered but it decreases to 81 when it also requires at least 6 hours of presence in that country).
- Little variations on the threshold that controls the second condition (static or dynamic context) affects in a very considerable way to the context detection (e.g. variations of only one per cent means over one hundred of difference in the number of contexts found, see Appendix B).
- Once a cluster has passed the last condition (“On the move” condition) there are three requirements that this cluster has to fulfil all together to be marked as “office” context. These conditions are based on the analysis if the following variables:
    - Time spent during weekends.
    - Time spent at night from Monday to Friday.
    - Amount of time spent during working hours.

  From these three thresholds, the strongest and more restrictive condition is the one that considers the amount of time spent on a cluster during working hours. The number of contexts identified does not change at all with any variation over this parameter as it is obvious (the clusters that pass the previous condition they just will change their status from home into office or vice versa), but as the share of time on a cluster during working hours required decreases, the number of offices detected raises up (not in a very intense way but still noticeable).

- It is possible to check that the sandwich condition has a very important effect in the context detection as well. If this number is kept stable, variations on the other parameters do not affect at some point (even if these variations are very big). It establishes the limit of contexts detected.

After these initial trials and with all the knowledge that the results have provided, a second stage with combinations of all the parameters takes place. The aim of the second set of tests is to determine the most appropriate values for the parameters in order to use the results in a case example.

4.3 Case Example

In this section, one application of the results is described in detail. With the dataset explained above, two additional files containing questionnaires with demographic and presence information and other with service and application usage for the same users, a complete analysis of the end-user's service usage is carried out. In the first sub section, the final stage of the sensitivity analysis is explained. The aim is to choose a proper result file to test the effectiveness of the algorithm.

4.3.1 Data Preparation

The sensitivity analysis can be considered as the first step in the testing of the algorithm. In the second step, a set of parameters is selected to provide the results that will be used in a comparative study by using demographics. The idea is to verify the algorithm's effectiveness by comparing the number of end-users employed and the number of office contexts identified.

The file with the demographic information shows the next conclusions,

Occupation	Number of People
Employed	349
Student	92
Unemployed	28
Retired	28
Housewife	10
Invalid	2
Other	12
Total =	521

Table 2 – Demographic Information of the End-users Analyzed

Using this information, it is possible to test the accuracy of the algorithm by comparing the number of offices detected and the occupation for the same end-users.

The value of the parameters for the first trial are presented in the following table,

Sandwich Restriction	40 repetitions
On the Move condition	<4%
Weekend time condition	<10%
Working Hours condition	>65%
Night time condition	<10%

Table 3 – Values for the algorithm's parameters in the first trial

In the first attempt the goal is not only to choose "logical" values that raise the number of offices detected (the key aspect of the testing process) but also to balance between the trades-offs. From the questionnaire files it is possible to get demographic information for 490 of the 578 users analyzed (there is no demographic information for all the users because e.g. some of them do not answer to questionnaires or some others that answer have been excluded by the algorithm because they do not have more than three weeks of data).

The next table shows the demographic classification extracted from the questionnaires for the 194 users with at least one office context identified by the algorithm,

Context Detection	End-users
Perfect Match - Office Detected / Employee	127
Office Context Detected for Student	36
Office Context Detected - No Demographic Info. To Check	25
Office Context Detected for Unemployed	3
Office Context Detected for Housewife	1
Office Context Detected for "Other" status	2
Total End-users	194

Table 4 – Demographic matches for the users with office context identified 1

One important factor to consider is that the definition of "office" context used by the algorithm fits with the concept of a student that goes every day to the University and stays there for the approximately same share of time than an employee spends at work.

The accuracy and effectiveness of the algorithm can be analyzed with the resultant values:

127 matches + 36 students = 163 contexts properly detected → 163/194 = 84% of success in the contexts identified by the algorithm. This is a relatively high effectiveness taking into consideration the fact that 25 of these users cannot be checked and are out of the calculation of this value.

The value decreases considerably when comparing with the data obtained from the demographic information regarding the users analyzed:

127 Offices detected / 331Employees = 38%

The results guide to two main conclusions:

- The number of offices found by the algorithm is low considering the number of employees for which there is no "office" context identified.

- The algorithm detects "office" context for employees with high accuracy, so although the number of "office" contexts found is low, at least the offices identified are properly detected.

The algorithm should increase the number of offices detected what can be easily done by simply raising the number of sandwiches needed (sandwich repetition condition). In a second and last trial, the algorithm is run with the following set of values,

Sandwich Restriction	100 repetitions
On the Move condition	<4%
Weekend time condition	<10%
Working Hours condition	>65%
Night time condition	<10%

Table 5 – Values for the algorithm's parameters in the second trial

All the values remain in the second trial but the number of sandwiches needed to identify clusters, which has been raised to one hundred. This way, the number of "office" contexts rises to 213 (20 more than in the previous trial).

The results for these values can be shown in the following table,

Context Detection	End-users
Perfect Match - Office Detected / Employee	138
Office Context Detected for Student	37
Office Context Detected - No Demographic Info. To Check	31
Office Context Detected for Unemployed	3
Office Context Detected for Retired	1
Office Context Detected for Housewife	1
Office Context Detected for "Other" status	2
Total End-users	213

Table 6 – Demographic matches for the users with office context identified 2

The algorithm's effectiveness can be calculated with these data in the same way as in the first trial:

138 matches + 37 students identified with office / 213 = 82.2 %. Again, the users without demographic information are out of the calculation of this value (31 users). The results are very positive considering the fact that 31 of 213 users represent the 14%.

Comparing with the data from the demographics:

138 Offices detected / 331Employees = 42.3 %

This means a small improvement of more than a 4% in the effectiveness caused by the increase in the number of offices detected.

If the number of repetitions needed to detect clusters is increased over one hundred, the variations in the results are insignificant (see Appendix B). Because of this, the service usage study explained in the last section of this chapter is carried out by using the results of the algorithm for the second trial mentioned here.

4.3.2 Effectiveness of the Algorithm

There exists a more rigorous way of testing the algorithm's effectiveness than by using the demographic information mentioned above. With one of the files containing the answers of the end-users for the question *"Which is your context at this moment?"* (i.e. home, working/studying or on the move) a better analysis of how the algorithm works can be done. With the result files provided by the algorithm and considering that every answer of this questionnaire is marked with a timestamp, it is possible to compare for those moments which context identifies the algorithm and what is the answer of the end-user for the question mentioned.

The accuracy of the algorithm for the original set of 578 users is tested. Merging the files containing the mapping information (context identified in blocks of five minutes) and the answers of the questionnaires, the results obtained are shown in the following table:

Answer of the User	Context detected	Number of Cases
Home	Home	286
Working / Studying	Home	112
Other (e.g. downtown)	Home	35
Home	Office	10
Working / Studying	Office	40
Other (e.g. downtown)	Office	4
Home	On the Move	36
Working / Studying	On the Move	27
Other (e.g. downtown)	On the Move	32

Table 7 – Comparative analysis of questionnaires and algorithm's results for 578 users

The accuracy of the detection for "home", "*office*" and "on the move" contexts can be easily calculated as the number of matches for one specific context divided by the number of cases where this context was detected by the algorithm, including all the mistakes (e.g. cases where this context home is identified and the user answers working or moving). This way, the results are:

- Home:

  286 matches / 433 "home" contexts detected = 66% of accuracy in "home" context detection.

- Office:

  40 matches / 54 "office" contexts detected = 74% of accuracy in "office" context detection.

- On the Move:

  32 matches / 95 "on the move" contexts detected = 34% of accuracy in "on the move" context detection.

The global effectiveness can be calculated as all the matches divided by all the cases:

358 perfect matches / 582 cases = 51%

Some interesting conclusions regarding to these results can be extracted. Firstly, “on the move” context presents the worst rate of accuracy and, because of this, it decreases the global effectiveness of the algorithm. This proves that the biggest difficulty in the context identification is to establish the rule that says whether a cluster is static or not. Anyway, it is promising the big number of offices detected correctly, confirming the ideas exposed in the previous section with the demographic analysis: most of the offices detected by the algorithm are properly identified. Secondly, it is very interesting to check that the effectiveness in “home” context detection is decreased by the number of real offices detected as homes. In every one of these situations, the office is not marked as the biggest home but as a second household. It is important to remember that the context is divided into sub-contexts and for this reason a single user can present more than one “home” context. These sub-contexts are sorted regarding to time spent on them. So, if office is detected as another household, this means that the algorithm can be improved or simply that this user has not standard schedule and this is why this office has been detected incorrectly.

A thorough analysis of the result file with the mapping information has proved that there are many users with big holes of missing information (i.e. situations where the handset is switched off and it is turned on again some days after generating a significant lack of data). Despite this, the results obtained for users with no big holes are basically the same that the ones obtained for all the end-users so it is possible to generalize the values calculated here.

4.3.3 Service Usage Study

In this section, an application of the results generated with the set of parameters chosen in the previous point is described in detail. The aim is to use the correct matches (i.e. confirmed employees with the demographic files and users with “home” and “office” contexts detected) in a service usage study. For this purpose, new data files regarding to application and service usage in time and data traffic for the mentioned users are utilized together with the output files that the algorithm provides.

4.3.3.1 Presence on Context

Before the service usage analysis, the effectiveness of the algorithm can be interpreted with the results obtained by using illustrations of the time spent in every context. It is important to remember again the fact that in this and the following figures, only users with "office" context properly detected have been analyzed.

The amount of time spent per context calculated using the algorithm can be compared with real data extracted from Statistics Finland (see Smura 2008 and Statistics Finland 2001),

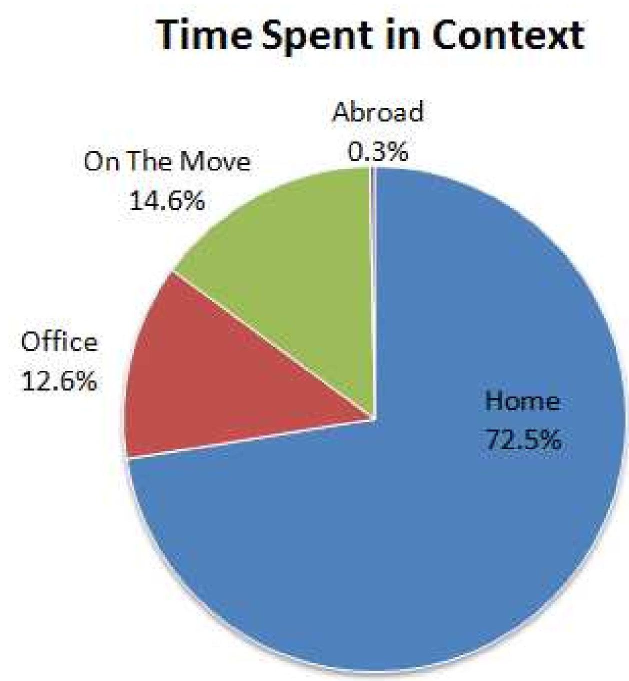


Figure 12 - Amount of time spent per context using the results of the algorithm

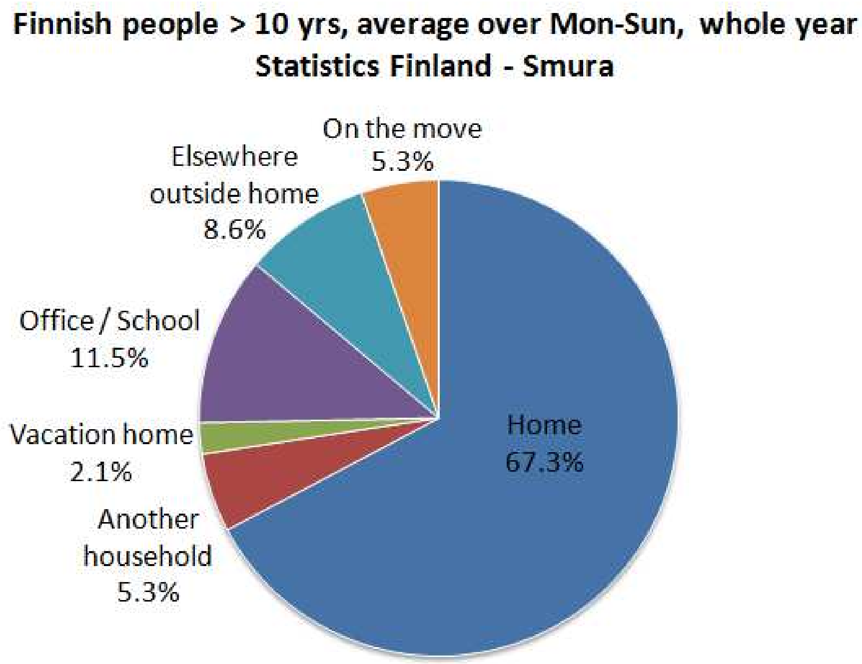


Figure 13 - Amount of time per context extracted from Statistics Finland

It is promising the proximity in the values of time spent at "home" (especially if the field "Another household" in the second figure is added, what means a 72.6% obtained from real statistics when a 72.5% was detected by using the algorithm) and "office" contexts (12.6% for the algorithm versus 11.5% obtained with real statistics). This means that the results of the algorithm fits with the real values extracted from statistics when excluding users which contexts have not been properly detected. The big difference appears with the values of the time spent "on the move" for both graphs, being considerably bigger the results obtained using the algorithm (14.6%) than the real values extracted from statistics of Finnish users. The reason of this difference is the fact that the algorithm does not detect all the possible contexts but only four of them. Besides, the sample used affects in a very significant way and it is important to remember that the dataset analyzed regards information of less than three months of data (from October 2007 to January 2008) including the winter holidays what can bias the results.

The share of presence of the users by hours of the day is another graph that provides valuable information. It can be easily checked how the presence at home decreases during working hours in the same way office presence increases at the same time through this figure. But the most meaningful fact is the peak of presence on the move from two o'clock p.m. to six o'clock in the afternoon, the time when most of the working users leave work and they have their leisure time. Curiously, there is no similar peak in the morning when users are supposed to go to work. It can be explained by the randomly chosen plans and paths coming up at the end of a working day.

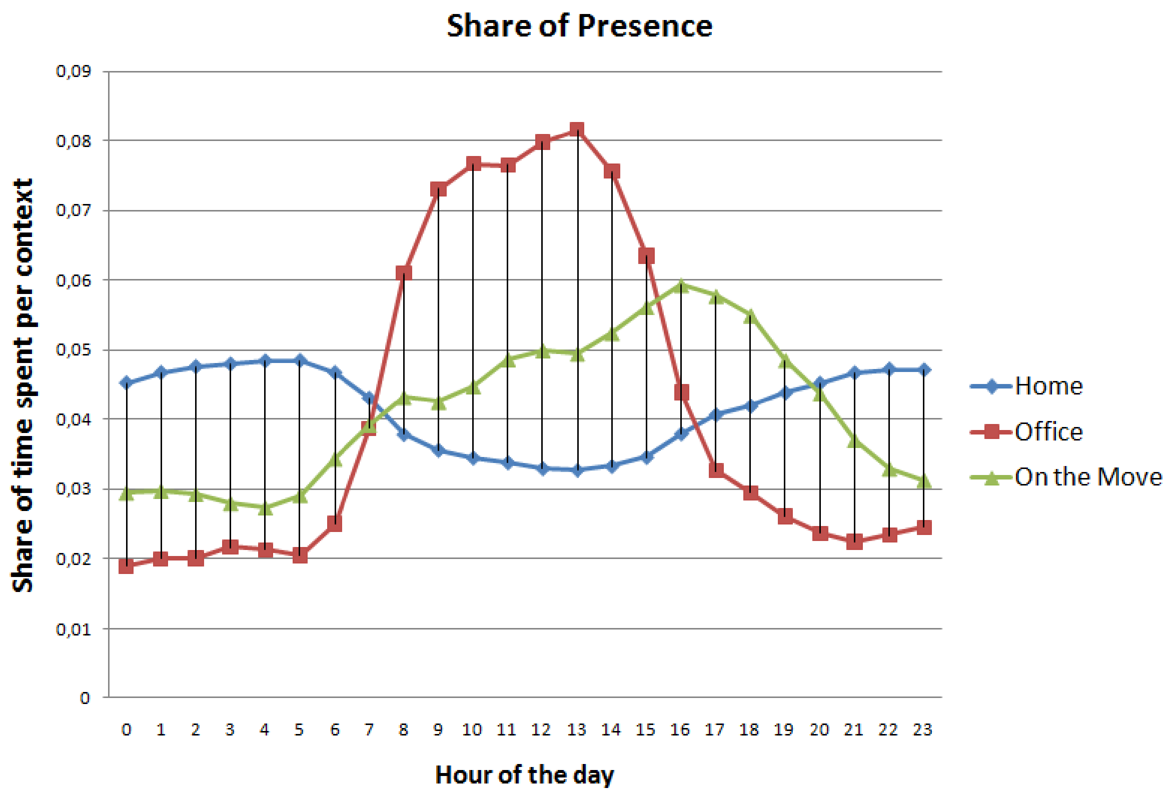


Figure 14 - Share of time per context per hour of the day

Additional charts comparing statistics and algorithm results are shown in the Appendix C.

In the following sections the service usage analysis is carried out. With the information of the output files of the algorithm and the files of application and service usage regarding time and traffic generated for the same users, research under a contextual point of view can be elaborated.

4.3.3.2 Service Usage

In the first figure, the total mobile service usage regarding time for some of the most used applications is shown,

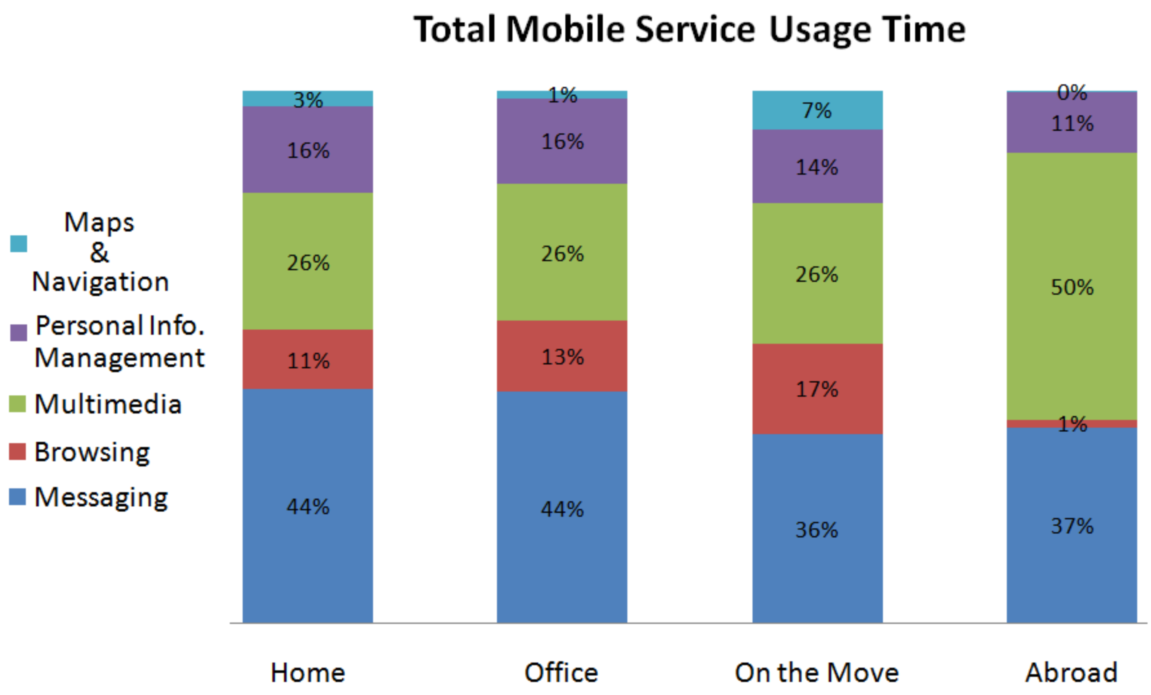


Figure 15 - Total mobile service usage across contexts

Although "*Maps & Navigation*" are used more at "on the move" context (for obvious reasons the needs of location information on the move are higher) the main conclusion analyzing time is that applications are used similarly in all contexts.

The following figure is particularly interesting. Although most of user's presence is at "home" context as it has been seen before, this is not the most active context. The next figure illustrates this finding, were the intensity of usage for every application in every context has been calculated. The idea is to divide the usage time by the time spent on every context (for the period of time studied). The night time (from 23 p.m. to 7 a.m.) has been omitted because the usage is very low during sleeping hours and the user can be considered as inactive at these moments.

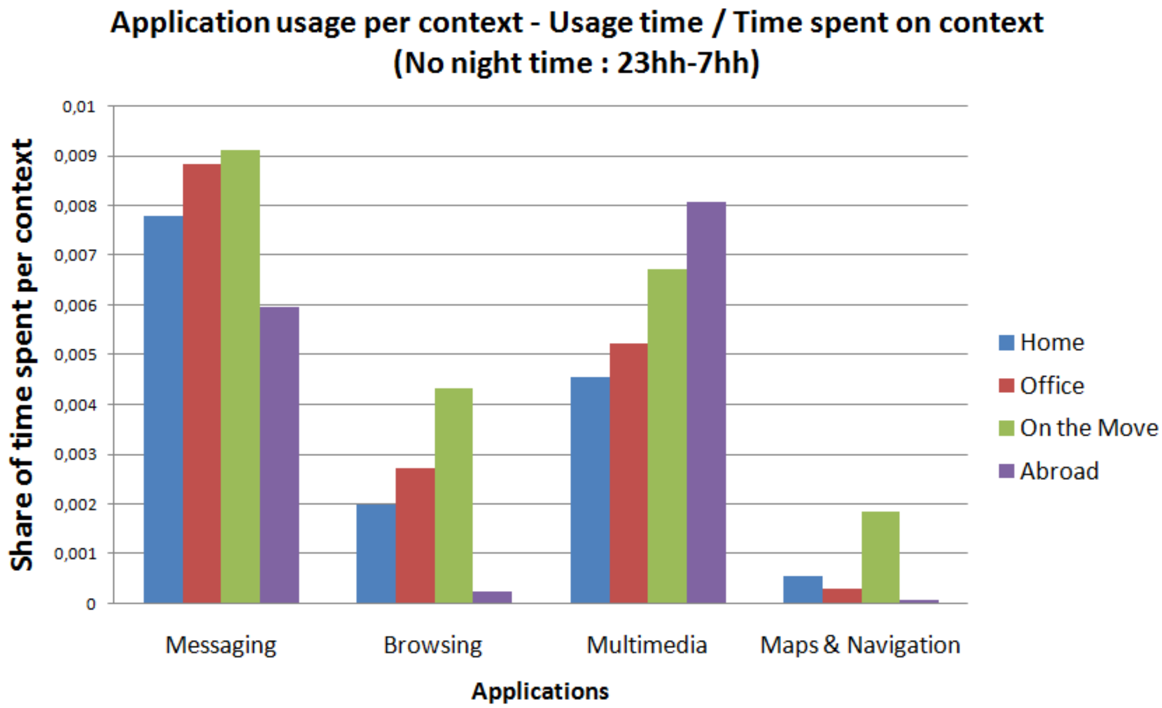


Figure 16 - Intensity of usage across contexts

The figure shows how "on the move" context is the most active in every application (especially in *Browsing*, *Multimedia* and *Maps & Navigation*) closely followed by "office" context. "Home", where the user spends most of the time, is the less active context.

It can be meaningful to analyze how the end-user spends his time by checking the average usage session duration. From figures like this, it is easy to get interesting conclusions as the longer usage of browsing and games (e.g. the game "*Snake*") at "on the move" context. Again, the longer usages correspond to "on the move" and "office" instead of "home".

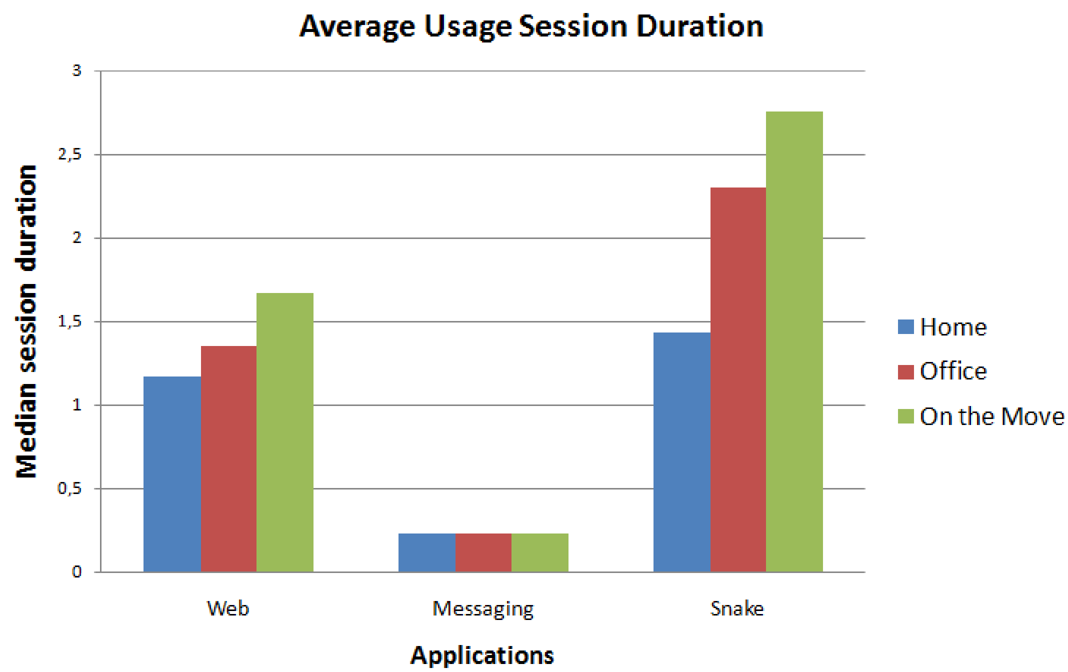


Figure 17 - Average usage session duration

4.3.3.3 Data Usage

The time analysis is not the only possibility available. Studies considering data under the application or the access technology point of view can be easily done by using the same files. The total amount of data generated per context is shown in the next figure,

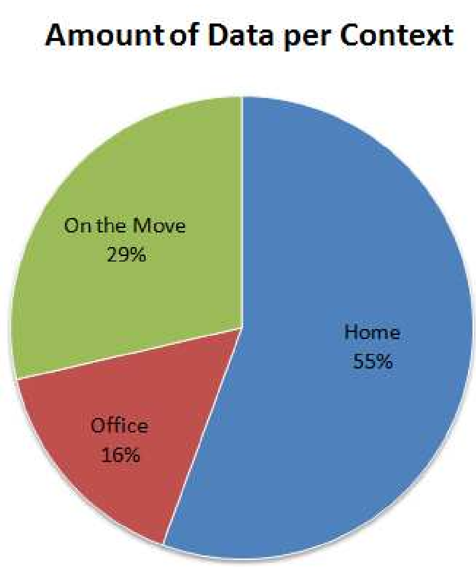


Figure 18 - Amount of data generated per context

It is at “home” were the biggest amount of data is generated (55%), followed by “on the move” (with a very representative 29%).

Taking into consideration the total amount of traffic generated per application exclusively the conclusions are that messaging usage decreases at “on the move” while browsing increases comparing this to “home” or “office” (both have a similar distribution). The following figure shows this,

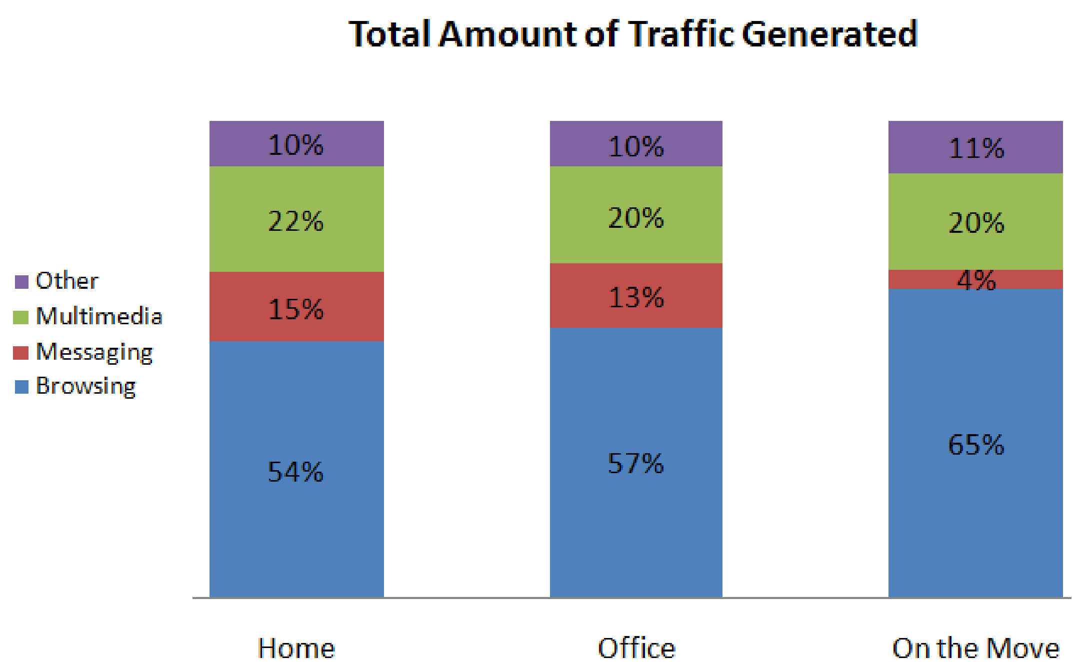


Figure 19 - Total amount of traffic generated per application per context

In the last two figures, a study considering the access technology used has been carried out. This is just another example of what can be done with the output files of the algorithm: adoption of new technologies analysis regarding context.

In the following figure, the total amount of data generated with all the possible technologies has been represented. If WCDMA is the dominating access technology (especially at “office” context) it is interesting to see how the most usage of WLAN is at “home”,

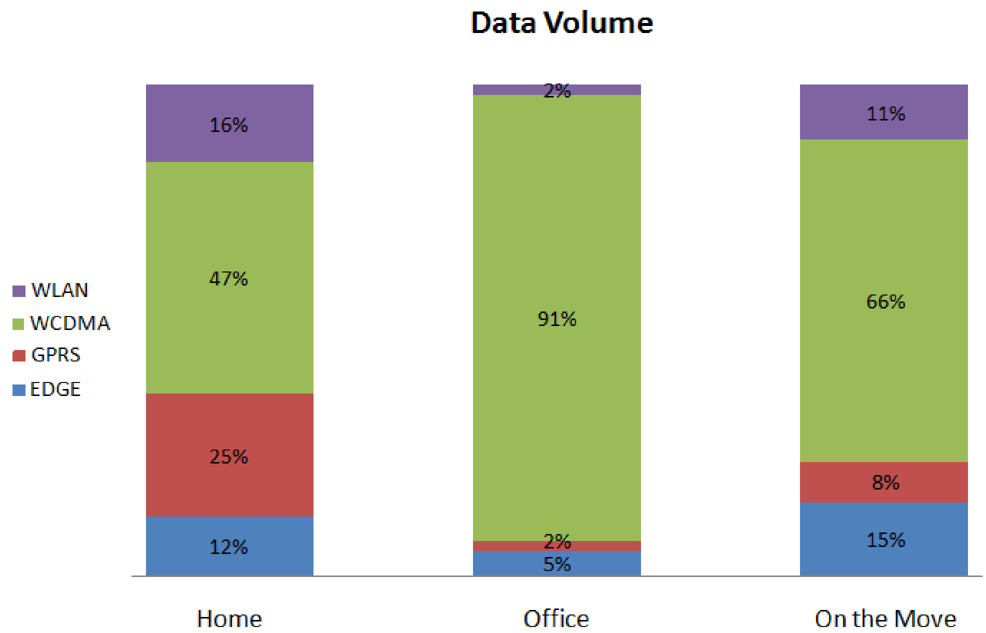


Figure 20 - Data volume per access technology per context

The last figure, where time instead data is analyzed, corroborates the conclusions extracted from the previous one giving new information: while WCDMA is equally distributed under time conception in the three main contexts, more than 60% of WLAN usage is at home,

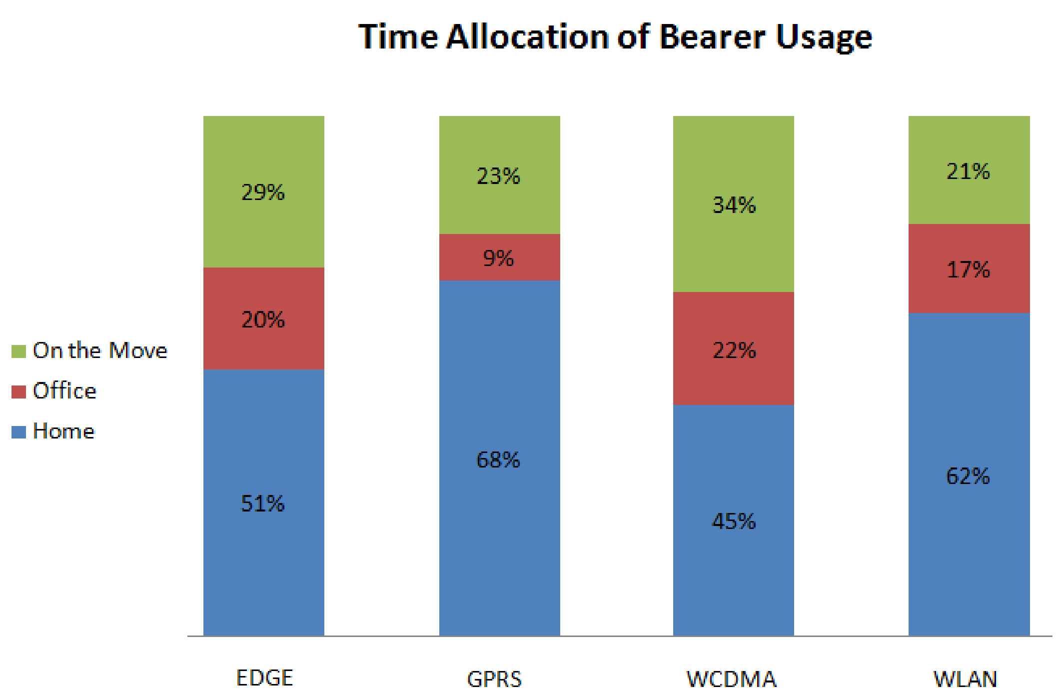


Figure 21 - Time allocation of bearer usage per context

These are a few examples of the possible applications of the results of the algorithm. Furthermore, the algorithm allows the analysis of service usage, adoption of technologies among many others under a new perspective: the end-user context.

4.4 Discussion

From the analysis of end-users service usage described in the previous section and elaborated by using proper identifications (i.e. employees with “office” and “home” context detected) important conclusions can be extracted. First of all, a basic analysis of usage of applications considering time exclusively proves that most of the presence of mobile end-users is obviously at home. Nevertheless, intensity analysis taking into consideration the time spent on every context have shown that “on the move” context is the most active, followed by “office” in every one of the applications examined (particularly in *Browsing*, *Multimedia* and *Maps & Navigation* utilities). The intensity of usage abroad is most active in *Multimedia* and *Messaging* what fits with the logic assumption that when the users are out of the country they prefer *Messaging* to communicate and *Multimedia* is the most used application (what include games, multimedia and infotainment).

The share of presence has shown the logic relation in the decrease of presence at home when the share of presence at office rises at the same time. On the other hand, it is curious the conclusion of the analysis of presence “on the move”. The figure 14 reflects a peak of presence “on the move” from 14 p.m. to 18 p.m. when the users leave “office” but there is no peak in the morning when users go to office. The most probable reasons of this behaviour are the fact that in the morning, the movement is more concentrated but the time to leave office is more spread along the afternoon, when the users move more e.g. downtown in that leisure time.

Regarding data analysis, WCDMA has been proven as the most used access technology at “office” context (over a 91% of the total data volume has been generated using this technology at office context) whereas most of the usage of WLAN (more than a 60% of the time usage at home) takes place at "home" (the needs for mobile Internet connection at home are lower once a standard user has, at least, one computer).

Finally, and considering the algorithm exclusively, it is important to mention the challenges in identifying "office" context for everybody. While the effectiveness of the algorithm is very high in the sense that most of the contexts are properly identified, there are a number of employees for whom "office" context was not detected. Although the algorithm can be still improved there are some reasons that make the task of office detection something difficult even if the algorithm is very well programmed. The main causes of this limitation are:

- There is less presence at "office" than in the other contexts. This is motivated by the fact that some users move quite often while working, they do not have one single place or their timetable is not standard (i.e. working hours out of the range from 8 a.m. to 18 p.m.).
- There is a challenge of subjectivity in generalizing the results (e.g. the difficulties in defining what is working hour or night time).
- The intrinsic problems regarding data processing: alteration of the results by outliers. The data used in this study concerns mobile end-users with less than three months of information including winter holydays what can considerably affect to the detection. The better the dataset used (probably a bigger sample without vacations) the better results.
- The impossibility of testing the algorithm because of the lack of demographic information or questionnaires to compare the results with.

5 Conclusions

5.1 Findings

The main findings of this thesis are the demonstration of the possibility to extract contextual information from the datasets available and the application of the contextual results into service usage analysis.

The analysis of the effectiveness of the algorithm has shown the accuracy in context detection when comparing the results to questionnaires. Additionally, comparisons show how close the output of the algorithm is to official European statistics even when the datasets present some problems.

The identification of context can provide a new dimension to studies related to service usage or technology adoption, for instance to the study of competition between emerging access technologies (see Smura 2006). The study shows that WCDMA is the most used technology and that WLAN is mostly used at “home”. On the other hand the service usage analysis proves that intensity figures can provide more accurate information than absolute total usage analysis. For instance, even when “home” is the most frequent context, it is not the most active. “On the move” and “office” contexts are more active for the most used applications (*Messaging*, *Browsing* and *Multimedia*).

5.2 Exploitation of Results

The process of context extraction, even without geographical information associated, has an important business value. The possibility of supporting targeted marketing applications with a better segmentation of users is one of the possible track of exploitation (see Uronen 2008).

New services related to social networking (e.g. a tool able to inform the users of when a friend is in a close context, see Eagle 2005) or more focused real-time marketing (as companies that provide service to shops in the way that when a user is close to this shop, it can broadcast sales information) are other examples of exploitation.

Gradually improving versions of context and location information extraction can be the beginning of new services and applications that together with the mobile Internet will greatly influence the telecommunications market evolution.

5.3 Limitations and Future Research

The research shows the difficulty of the “office” context identification. “Office” context has been only detected for one third of the end-users analyzed (most of them are employees and students). The source of this “office” challenge is the inherent nature of the algorithm. The rules used to detect offices are very general because they have to be valid for most of the population under analysis. The subjectivity in the thresholds and the rules themselves are one of the most important limitations of the thesis.

On the other hand a more detailed context analysis can provide significant new information. To know whether the end-user is “driving” or “walking” while on the move (understanding driving and walking as “on the move” sub-contexts) is another potential improvement. Once the accuracy has been checked and improved, a second step regarding sub-contexts can be taken. New and more precise rules for expanding the identification to more detailed contexts are the basic challenge for further research.

In addition, the integration of a context identification service in the handsets (through the development of an integrated application) is another interesting point. The adaptation of the devices to the contexts (e.g. while the user is at office context, the handset turns into silent profile by itself) is full of advantages. Further research on how to integrate the algorithm in the logic of the device, and how to process the information just in time are the main questions to be discussed. For practical reasons, the application developed along this research has not been focused on the “real time” processing aspect. The implementation has been designed without considering how to minimize the executing time, how to send the reports to the data processors (if the logic is not integrated in the device) or how to create a consistent code to be included as an application in the logic of the handsets.

Other lines of investigation on detecting end-user movements considering mobile handset’s signals are open to create different and more accurate modeling contexts techniques (e.g. using jitter measurement techniques and other parameters, see Kawashima et al., 2006).

For future research, more accurate datasets to work with become necessary. This was a major difficulty in the programming and testing stages. The raw data suffered from missing information (some users presented holes in their log files with more than three days without any report) and values in a wrong format that made them useless (like a wrong timestamp). Most of the problems are a result of a dataset with winter holidays in between, information of a short period of time (less than three months), a lack of demographic information and the absence of more questionnaires regarding contextual situations. If no better datasets can be obtained, special situations (e.g. holidays, holes in the data, users with few information, etc) should better be controlled in order to get more reliable results.

But, without any doubt, the objectives in the short term are the improvement of the algorithm with a more detailed level of sub-context analysis, more parameters besides time and the usage of different data mining approaches.

6 References

Alpaydin, E. 2004. Introduction to Machine Learning (Adaptive Computation and Machine Learning). MIT Press.

Balakrishnan, V. K. 1997. Graph Theory, McGraw-Hill; 1st edition. February, 1997.

Bazire, M. & Brézillon, P. 2005. Understanding Context Before Using It. Modeling and Using Context. Chapter 3:29-40. Springer Berlin / Heidelberg.

Barabási, A.L. 2003. Linked. A Plume Book.

Bauer, J. M. 2005. Bundling, Differentiation, Alliances and Mergers: Convergence Strategies in U.S. Communication Markets. Michigan State University. Communications & Strategies, No. 60, p. 59, 2005.

Bernardos, C. & Soto, I. & Calderón, M. 2005. IpV6 Network Mobility. http://www.cisco.com/web/about/ac123/ac147/archived_issues/ipj_10-2/102_ipv6.html. Referred 06.05.2008

Berson, A. & Smith, S. & Thearling, K. 2000. Building Data Mining Applications for CRM. McGraw-Hill.

Best, J. 2006. Yahoo breaking the walled garden. http://networks.silicon.com/mobile/0,39024665,39160005,00.htm Referred 15.01.2008.

Bishop, C. M. 2006. Pattern Recognition and Machine Learning, Springer.

Busygin, S. & Prokopyev, O. & Pardalos, P. 2008. Biclustering in Data Mining. Computers & OR 35(9): 2964-2987.

Caldwell, C. 1995. An Interactive Introduction to Graph Theory. http://www.utm.edu/cgi-bin/caldwell/tutor/departments/math/graph/intro. Referred 06.05.2008.

Carringhton, PJ. & Scott, J. & Wasserman, S. 2005. Models and Methods in Social Networks Analysis. Cambridge University Press.

Chau, F. 2007. Breaking down the wall: Cellcos are increasingly recognizing that the “walled garden” content paradigm simply won´t play in the new IP world and they will have to open up their networks to third parties. But to stay profitable in this new open business environment, cellcos need to be creative and pay close attention to end end-user demands. http://findarticles.com/p/articles/mi_m0FGI/is_1_18/ai_n19041152. January 2007. Referred 15.01.2008.

Chen, G. & Kotz, D. 2001. A survey of context-aware mobile computer research. (No TR200-381). Hanover, Dartmouth College, Department of Computer Science.

De Nooy, W. & Mrvar, A. & Batageli, V. 2005. Exploratory Social Network Analysis with Pajek. Cambridge University Press.

De Veaux, R. 1997. Book review of Neural Networks for Pattern Recognition and Pattern Recognition and Neural Networks, International Journal of Neural Systems, 8,2, p.249-250. Department of Mathematics and Statistics, Williams College (1997).

Dietterich, T. G. 1999. Machine Learning. In Rob Wilson and Frank Keil (Eds.) The MIT Encyclopedia of the Cognitive Sciences, MIT Press: 497-498.

Eagle, N. 2005. Machine Perception and Learning of Complex Social Systems. Doctoral dissertation, Massachusetts Institute of Technology.

Eiek, S. T. 1996. Aspects of Network Visualization. At & T Laboratories. Computer Graphics and Applications, IEEE. Volume 16:69-72. USA.

Esbjörnsson, M & Weilenmann, A. 2005. Mobile Phone Talk in Context. In Proceedings of Context'2005 – The 5th International and Interdisciplinary Conference Modeling and Using Context. Springer Verlag, 11:140-154.

Fields, A. 2005. Discovering Statistics Using SPSS. Sage Publications.

Gross, T. & Specht, M. 2001. Awareness in context aware information systems. Mensch & Computer, Germany, March, 2001.

Harmonised European Time Use Survey, 18.07.2007 https://www.testh2.scb.se/tus/tus/Graph_0.html Referred 20.05.2008

IDC Vendors, 2007. The Business Value of Social Networking Applications. Adapted from Happe, R. U.S. Social Networking Application 2007–2012 Forecast and Analysis. http://www.hivelive.com/files/hl_idc_spotlight.pdf . Referred 06.05.2008.

ISO 13407, 1999. Human-centred design processes for interactive systems. International Standard, the International Organization for Standarization.

Jahn, O. & Möhring, R. & Schulz, A, Moses, N. 2005. System-optimal routing of traffic flows with end-user constraints in networks with congestion. Institute for Operations Research and the Management Sciences (INFORMS), Linthicum, Maryland, USA, 2005.

Jorstad, I. & Dustdar, S. & Thanh, D. 2004. Evolution of Mobile Services: An Analysis of Current Architectures with Prospect to Future. L. Baresi et al. (Eds): UMICS 2004, LNCS 3272, pp. 125-137, 2004. Springer-Verlag Berlin Heidelberg 2004.

Kaasinen, E. (2003). End-user needs for location-aware mobile services. Personal and Ubiquitous Computing, Volume 7, Number 1, May 2003. 70-79.

Karp, S. 2008. Creating Customized Social Networking Applications for Business. http://publishing2.com/2008/02/22/creating-customized-social-networking-applications-for-business/. Referred 06.05.2008.

Kawashima, Y., Mineno, H., Ishihara, S. & Mizuno, T. 2006. Evaluation of method for multiple path distribution based on delay-jitter. SAINT Workshops 2006: 109-112.
Kivi A. 2007. Measuring Mobile End-user Behavior and Service Usage: Methods, Measurements Points, and Future Outlook. Proceedings of the 6th Global Mobility Roundtable, 1-2 June 2007, Los Angeles, California, U.S.

Krieger, C .(1996). Neural Networks in Data Mining. http://www.cs.uml.edu/~ckrieger/end-user/Neural_Networks.pdf . Referred 24.04.2008.

Laakso, S., Krings, M. & Kuhl, M. 2008. Nokia Mobile Context Data Analysis. Collaborative Innovation Networks (COINs). Joint course offered by University of Cologne, Helsinki University of Technology and University of Salento. January 2008.

Laasonen, K. & Raento, M. & Toivoinen, H. 2004. Adaptative On-Device Location Recognition. Second International Conference on Pervasive Computing (Vienna), April 23, 2004.

Lee, I., Kim, J. & Kim, J. 2005. Use Contexts for the Mobile Internet: A longitudinal Study Monitoring Actual Use of Mobile Internet Services. International Journal of Human-Computer interaction 18(3), 269-292, 2005.

Li, F. Li & Whalley, J. 2002. Deconstruction of the telecommunications industry: from value chain to value networks, Telecommunications Policy 26 (2002), pp. 451–472.

Liebowitz, S. & Margolis, S. 1996. Network Externalities Effects. http://www.utdallas.edu/~liebowit/palgrave/network.html . Referred 06.05.2008.

Llora, J. & Garrell, J. M. 2001. Evolution Of Decision Trees. Research Group of Intelligent Systems. Ramon Llull University. Barcelona, Spain, 2001.

Long, S., Aust, D., Abowd, G. D., and Atkeson, C. G. 1996. Rapid prototyping of mobile context- aware applications: The cyberguide case study. Proceedings of the 2nd annual international conference on Mobile computing and networking, 97-107. New York, United States.

Newman, M. 2007. Is the Internet a telco slayer or telecoms saviour? http://www.informatm.com/itmgcontent/icoms/s/sectors/mobilestrategies/20017441515.html 13th of July 2007. Referred 10.01.2007

Nilsson, J. N. 1996. Introduction to Machine Learning. An early draft of a proposed text book. Robotic Laboratory Department of Computer Science, Stanford University. Stanford 1996.

Olafsson, S. & Li, X. & Wu, S. 2006. Operations research and data mining. Department of Industrial and Manufacturing Systems Engineering, Iowa State University. USA, November 2006.

Padovitz, A. & Loke, S. & Zaslavsky, A. & Burg, B. & Bartolini, C. 2005. An Approach to Data Fusion for Context ananess. Fifth International Conference on Modeling and Using Context, CONTEXT'05, Paris, France, July 2005.

Peltomäki, M. & Alava, M. 2008. Correlations in Bipartite Collaboration Networks. Laboratory of Physics, Helsinki University of Technology. Finland, February 2008.

Raento, M. & Oulasvirta, A. & Petit, R. & Toivonen, H. 2005. ContextPhone - A prototyping platform for context-aware mobile applications. IEEE Pervasive Computing, 4 (2): 51-59. 2005.

Rohlf, T. & Winkler, T. 2008. Network Structure and Dynamics, and Emergence of Robustness by Stabilizing Selection in an Artificial Genome. USA, May 2008.

Ryszard S. Michalski, Jaime G. Carbonell, Tom M. Mitchell (1983), Machine Learning: An Artificial Intelligence Approach, Tioga Publishing Company.

Schilit, B.N. & Adams, N.J. & Want, R. 1994. Context-Aware Computing Applications. In Proceedings of the Workshop on Mobile Computing Systems and Applications (IEEE Computer Society). Santa Cruz, CA. pp: 85-90, 1994.

Schmidt, A. & Beigl, M. & Gellersen, H.W. 1998. There is more to Context than Location. In Proceedings of Workshop on Interactive Applications of Mobile Computing. Rostock, Germany, November 1998.

Schonfled, E. 2007. Radar turns mobile pictures into conversation starters. http://www.techcrunch.com/2007/11/05/radar-turns-mobile-pictures-into-conversation-starters/ 5th November 2007. Referred 7.01.2007.

T. Smura, 2006. Competition between Emerging Wireless Network Technologies: Case HSPA vs. WiMAX in Europe, in 17th European Regional ITS Conference, August 22-24, 2006, Amsterdam, Netherlands, 2006.

Smura, T. 2008. Access alternatives to mobile services and content: a techno-economic analysis, in ITS 17th Biennial Conference, June 24-27, 2008, Montreal, Canada.

Statistics Finland, 2001. Time use survey. Available at: http://www.stat.fi/meta/til/akay_en.html . Referred 20.05.2008.

Thearling, K. 2007. An introduction to Data Mining: Discovering hidden value in your data warehouse. White paper. http://www.thearling.com/text/dmwhite/dmwhite.htm Referred 11.4.2008.

Tools, 1, 2008. JGraph. http://www.touchgraph.com/ . Referred 20.02.2008.

Tools, 2, 2008. GraphViz. http://www.graphviz.org/ . Referred 20.02.2008.

Tools, 3, 2008. CondorView. http://www.galaxyadvisors.com/ . Referred 20.02.2008.

Tools, 4, 2008. Pajek. http://vlado.fmf.uni-lj.si/pub/networks/pajek/ . Referred 20.02.2008.

Tools, 5, 2008. InFlow. http://www.orgnet.com/ . Referred 20.02.2008.

Tools, 6, 2008. TouchGraph. http://www.touchgraph.com/ . Referred 20.02.2008.

Uronen, M. 2008. Market Segmentation Approaches in the Mobile Service Market. Master's Thesis, Networking Laboratory, Helsinki University of Technology.

Verkasalo, H. 2005. Handset-Based Monitoring of Mobile Customer Behavior. Master's Thesis, Networking Laboratory, Helsinki University of Technology.

Verkasalo, H & Hämäinen, H. 2006. Handset-Based Monitoring of Mobile Subscribers. Department of Electrical and Telecommunications Engineering. Helsinki Mobility Roundtable. Helsinki School of Economics.

Verkasalo, H. 2007a. A Cross-Country Comparison of Mobile Service and Handset Usage. M.Sc. Thesis, Helsinki University of Technology, February 2007.

Verkasalo, H. 2007b. Contextual Patterns in Mobile Service Usage. Networking Laboratory. Department of Electrical and Telecommunications Engineering. Working Paper. Helsinki University of Technology.

Verkasalo, H. 2007c. Measurement of Smartphone Service Evolution in Finland. Journal of Targeting, Measurement and Analysis in Marketing; Special Issue; December 2007.

Wasserman, S. & Faust, K. 1994. Social Network Analysis. Methods and Applications. Cambridge University Press. 1994.

WinterBottom, D. 2007. Social networking has potential to drive mobile revenues, but also to add to network woes. http://www.informatm.com/itmgcontent/icoms/s/sectors/mobile-content-apps/20017480656.html;jsessionid=3DAD42964D2B7E70EA1ECE0027565DD3 19th of November 2007. Referred 3.01.2008

Xiao, Y. & Shen, Y. & Du, D. 2007. Wireless Network Security. Signals and Communications Technology. Springer. 2007.

7 Appendices

7.1 Appendix A – Logic of the Algorithm Developed

The algorithm developed to extract the contextual information from the datasets provided by Nokia has been explained along the chapters three and four. However, only the main ideas have been presented in these chapters and, for this reason, a deeper analysis of the algorithm's logic is needed for a better understanding.

As commented in the section 4.1, the java code programmed takes a text file with information about end-users and timestamps of every transition between base stations. The algorithm processes the data as it will be explained before, and it provides four result files:

- Cell classification: this file contains two columns for every user, the base station and the context identified for that base station.

- Contexts identified: this file contains two columns, one for the user's identifier and another with a code that informs of the contexts identified for that user, as it follows:

  0: No context detected

  1: Home context detected

  2: Office context detected

  3: Home and Office contexts detected

  4: Abroad context detected

  5: Home and Abroad contexts detected

  6: Office and Abroad contexts detected

  7: Home, Office and Abroad contexts detected

- Presence mapping: this file presents three columns for every user, the date, an index obtained through the division of the day in blocks of five minutes (this means 288 indices) showing which moment of the day correspond to that row and the context for that moment. This file can be easily generated once all the base stations have been detected.

The program generates three network files for every user as well (network topological files for the network visualization tool used, i.e. Pajek):

- Network file with all the information
- Network file with information of the working week excluding weekends
- Network file with only weekend information

Before executing the Java file, the input file has to be pre-processed. This file must be sorted by users, date and time and it cannot have missing or wrong values (e.g. dates in a wrong format). The input file must have five columns:

- End-user id
- Date
- Time
- Base station reached at that moment
- Mobile Country Code

The first thing that the code does is to check the number of end-users and after that to exclude those with less than three weeks of information. The big file is split into as many files as end-users. Consequently, the whole process will be faster. In the creation of these files, a new column showing the day of the week is added to the file. From this point on, the algorithm adds columns to the file in every action. Once with the weekdays, the program starts the data processing.

Next all abroad situations are detected. The code analyzes the most used mobile country code and marks it as the code of the country of residence. Then, it starts to check different country codes and compares them with a list of all the existing country codes. If the mobile country code is not in the list, the line is erased. If not, the algorithm counts the number of consecutive cases where that mobile country code appears and the total amount of time spent out of the country of residence. If there are more than ten lines and more than six hours of presence in that country, the base stations involved are marked as the same cluster and the context is marked as abroad.

After this, all the duplicated cases are deleted (i.e. consecutive cases with same base station index) because every line represents a transition between base stations so it has no sense a jump from one base station to itself.

In the following action the time spent in every transition is calculated and then the time spent in every base station. This is the information needed in the clustering process that goes next.

The grouping of cells into clusters has been explained in the chapters three and four. Basically, in this action the algorithm searches for “sandwich” appearances, it counts them and if the number of repetitions is higher than a pre-established threshold, the algorithm groups the clusters giving the same base station identifier to both of them (the one of the cluster with more time spent on it).

The input file is full of duplicated cases after the clustering process. Because of this, the next action is to delete again all the consecutive rows with the same base station index.

With all the clusters identified, the algorithm starts to calculate the times needed for the context detection:

- Total time spent in the cluster
- Time spent at night (from midnight to six a.m. from Monday to Friday)
- Time spent during working week
- Time spent in the cluster during weekends
- Time spent during working hours (from 8 a.m. to 18 p.m.)

The next action is to take every cluster and analyze the context. A new column with information about the context for every cluster ("home", "office", "on the move" or "abroad") is added to the file.

In the following step, the algorithm goes deeper in the process of context detection by identifying all the sub-contexts. As it was explained before, it is possible to find e.g. more than one "home" context for a single user. These "homes" will be sorted considering the total amount of time spent as the key variable. In these cases the context is "home" and the sub-context is the ranking of that cluster in the list of all the "home" contexts identified. This is a simple way to illustrate that the biggest cluster is the most relevant and, probably, the real home of the user.

With all this information, the program is ready to prepare the network files used by Pajek software to represent the context graphically. First of all, the algorithm creates two additional files for every user: the weekend information (by erasing the information of the rest of the week) and the working week information (without Saturday and Sunday information). Thank to this division the idea of "working" and "not working" (see figure 5) can be more clearly expressed.

The method "*Text2Pajek*" takes every file and generates one basic network file (the input for *Pajek* software) with all the clusters as nodes and the jumps between base stations as arrows (these jumps are obviously directed because the movements among base stations always follow a direction). The links are weighted and these weights are the number of jumps between two base stations (this way the most used paths can be easily detected).

In the last step, the program gives sizes and colours to the nodes. For giving sizes, the algorithm uses the amount of time spent (the size is calculated as the square root of the total amount of time spent on a cluster divided by the total amount of time of the whole panel).

As it has been mentioned, the simple use of a colour code can provide very valuable information. The algorithm gives colour to the nodes considering time aspects. The day is divided into four groups of hours:

- Blue group: hours from 0 p.m. to 6 a.m.
- Green group: hours from 6 a.m. to 12 p.m.
- Orange group: hours from 12 p.m. to 18 p.m.
- Red group: hours from 18 p.m. to 0 p.m.

The algorithm checks which cluster belongs to every group. For this purpose, the algorithm calculates the time the user spends in every group and so each cluster is painted with the corresponding colour (assigned by group).

7.2 Appendix B – Sensitivity Analysis

In the following table, the results of the sensitivity analysis explained in the fourth chapter are presented.

Several trials are run for every restriction to help in the understanding of how variations over the mentioned parameters affect to the global results. The conclusions and explanations of these results are described in the section 4.3.

In the last rows, combinations of the values that produce the most interesting results are analyzed.

	Number of Repetitions Sandwich	1st Threshold - Context - On the Move	2nd Threshold Time Spent on Weekends	3rd Threshold Time Spent During Working Hours	4th Threshold Time spent at Nights	HOME	OFFICE	ABROAD	TOTAL	HOME	OFFICE	ABROAD
Sandwich Restriction	5	0.05	0.1	0.75	0.05	578	100	93	1227	1015	100	112
	10	"	"	"	"	578	120	93	1294	1061	121	112
	20	"	"	"	"	578	135	93	1343	1094	137	112
	40	"	"	"	"	578	155	93	1506	1232	162	112
1st Threshold: "On the Move" Condition	5	0.04	0.1	0.75	0.05	578	111	93	1320	1094	114	112
	"	0.05	"	"	"	578	100	93	1227	1015	100	112
	"	0.1	"	"	"	578	68	93	1000	820	68	112
	"	0.15	"	"	"	578	43	93	890	735	43	112
	"	0.2	"	"	"	578	16	93	797	669	16	112
2nd Threshold: "Weekend time" Condition	5	0.04	0.1	0.75	0.05	578	111	93	1320	1094	114	112
	"	"	0.25	"	"	578	116	93	1320	1089	119	112
	"	"	0.35	"	"	578	116	93	1320	1089	119	112
	"	"	No Restriction	"	"	578	116	93	1320	1089	119	112
3rd Threshold "Working Hours" Condition	5	0.04	0.1	0.75	0.05	578	111	93	1320	1094	114	112
	"	"	"	0.6	"	578	142	93	1320	1062	146	112
	"	"	"	0.5	"	578	154	93	1320	1049	159	112
	"	"	"	No Restriction	"	578	167	93	1320	1033	175	112
4th Threshold "Night time" Condition	5	0.04	0.1	0.75	0.35	578	116	93	1320	1089	119	112
	"	"	"	"	0.15	578	116	93	1320	1089	119	112
	"	"	"	"	0.05	578	111	93	1320	1094	114	112
	"	"	"	"	No Restriction	578	116	93	1320	1089	119	112

Combination of Parameters ***Rising the number of Sandwiches***	30	0.04	0.1	0.65	0.1	578	188	74	1478	1168	198	81
	40	0.04	0.1	0.65	0.1	578	194	74	1479	1195	203	81
	60	0.04	0.1	0.65	0.1	578	205	74	1523	1227	215	81
	80	0.04	0.1	0.65	0.1	578	208	74	1537	1237	219	81
	100	0.04	0.1	0.65	0.1	578	213	74	1551	1246	224	81
	200	0.04	0.1	0.65	0.1	578	216	74	1593	1284	228	81

7.3 Appendix C – Share of Presence per Context

The following pictures show the differences in time spent for the user between real statistics (see Harmonised European Time Use Survey) and the results of the algorithm.

It is particularly interesting to analyze the shape of the lines that represent "home" and "office". For "office" context, the resemblance among the lines of the two figures is clear. Another detail is the fact that "office" presence does not reach 0% at any hour in both figures, being higher the value for the results of the algorithm (this means that the values of the conditions for context detection used in the algorithm can be changed to improve the results).

In the shape of "home" presence, there are two relevant findings that deserve special attention. First, the presence at home regarding to the algorithm's result does not reach 90% on its peak, while in the figure showing real statistics the highest value is close to 100% (algorithm's accuracy can be improved).

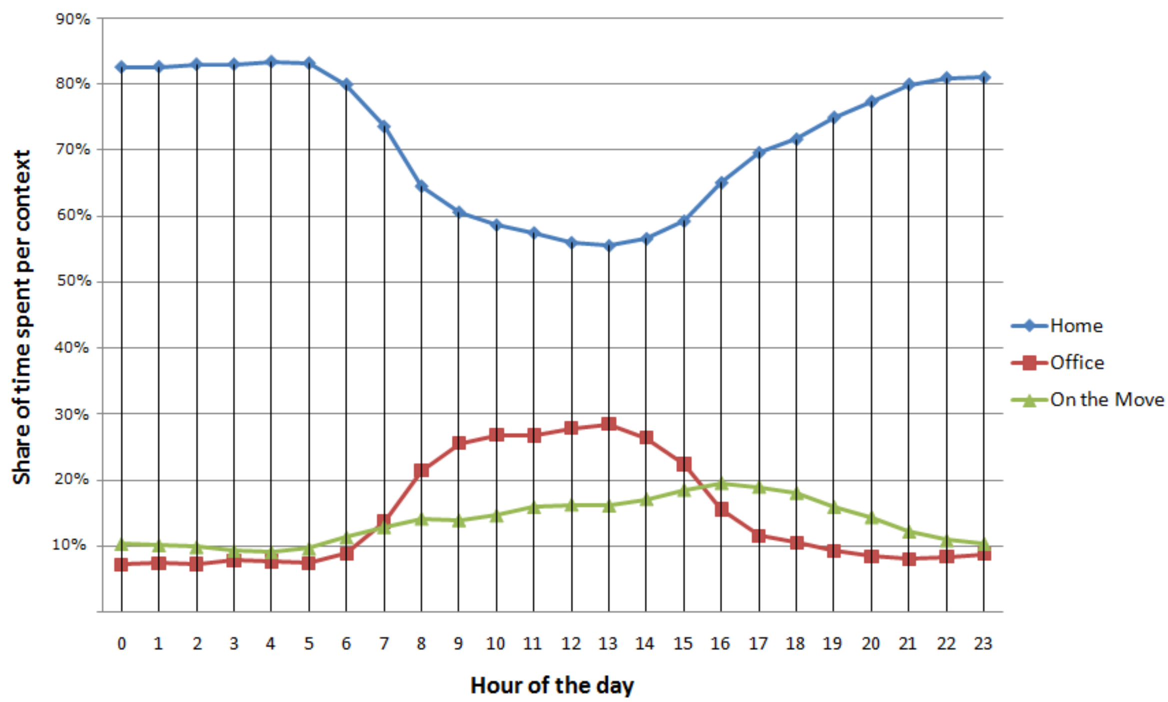


Figure 22 - Share of time per context obtained from algorithm results

Second, it is possible to check in both figures how "home" presence presents an intense fall during working hours. However, real statistics provides more information than the algorithm. In the figure 24, a second decrease from 14 p.m. to 21 p.m. can be seen. This has a logic explanation: end-users spend their leisure time while they are out of home or office and this time is longer than the one obtained with the algorithm results.

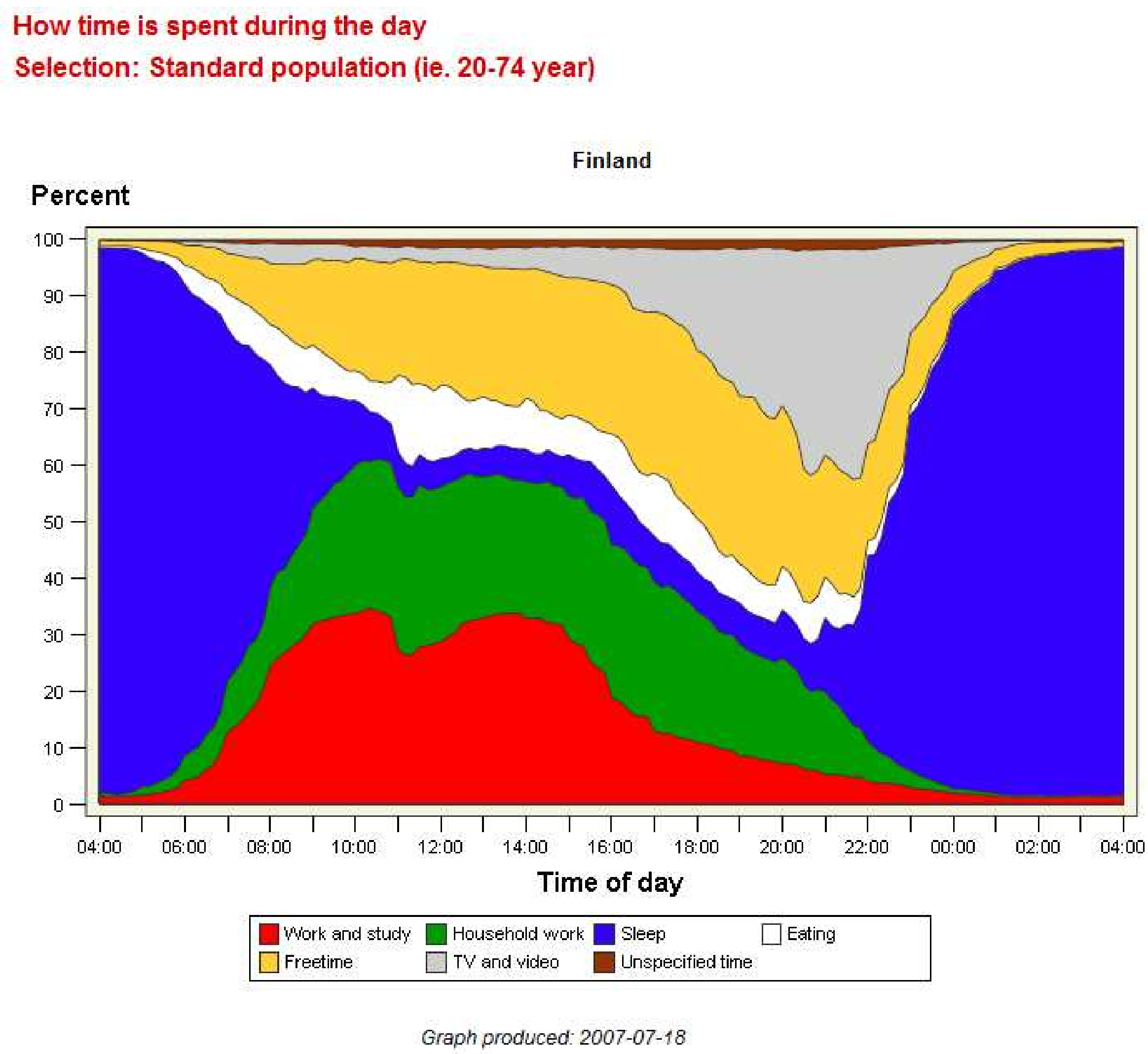


Figure 23 – Distribution of contexts during the day (see Harmonised European Time Use Survey)