UNIVERSITA' DEGLI STUDI DI PALERMO
DIPARTIMENTO DI INGEGNERIA CHIMICA, GESTIONALE, INFORMATICA, MECCANICA
Dottorato in Ingegneria Informatica

DESIGN METHODOLOGIES TO ENABLE COLLECTIVE INTELLIGENCE IN PLATFORMS FOR HUMAN-HUMAN INTERACTION

Settore scientifico disciplinare INF-INF/05

TESI DI
DARIO PIRRONE

COORDINATORE DEL DOTTORATO
PROF. SALVATORE GAGLIO

TUTOR
PROF. ROBERTO PIRRONE

XXIV CICLO ANNO ACCADEMICO 2012-2013

DOTTORATO

Contents

List of Figures

Chapter 1

Introduction

Over the past two decades, we have witnessed a huge change economic and social, which has affected the lives and activities of each of us. The introduction of the PC, and in particular of Internet, has been one of the engines of world economic growth within the Information and Communication Technology (ICT). Particularly the coming of web 2.0 (O'Reilly (2005)) has led to a real social revolution. Although the definition of Web 2.0 is an operative one, the term is often used to explain changes in the way users approach to web platforms. In a nutshell the term Web 2.0 refers to the set of technologies available on the web that facilitate information sharing, distance communication and collaboration.

Originally, the web is conceived as a way to display static hypertext documents (Web 1.0), lather we witness to birth of many web solutions that allow users to have an high degree of freedom in the web content production. Information management is no longer an exclusive prerequisite for computer specialists but becomes a public domain task. There is a growing desire to communicate and to share own thoughts and own personality on the network through the community. The companies provide increasingly web-oriented systems for access to their databases by sharing with users, allowing them to interact with the sources of information. These new of web systems type have the purpose of bringing together people interested in common topics and allow them to interact through chat or personal home-pages that favor the publication of contents created by the users.

The time is ripe for the birth of the first social networks sites, the best example of web 2.0 applications. In 2003 we are witnessing the debut of MySpace, which in 2005 records a number of visited pages higher than Google's ones. The year 2004 sees the Facebook coming, that borns in the Harvard University and then it developes rapidly growing, until to become the social networking site of most successful in the world. In the world of social networks, the more successful recent phenomenon is Twitter, born in 2006, but between 2008 and 2009 becomes, after Facebook and MySpace, the third most visited social networking site. It is

estimated that today there are over 200 social networking sites.

The Web is become an ecosystem of participation, where value is created by the aggregation of many individual user contributions. The need arises to exploit the potential of the aggregation of these information in order to generate new knowledge to share with all system users. These needs pose new challenges in the research field of knowledge acquisition and organization. New user generated contents are more and more complex and their aggregation has both economical and social importance. The way contents are re-aggregated and organized leads to new knowledge.

The work presented in this thesis is focused on the analysis of the processes responsible for the generation of Collective Intelligence (CI), i.e. the knowledge resulted from the collaboration of many individuals mediated by an intelligent system. In such systems, man and machine work together to produce Emergent Knowledge, which is the knowledge that is not directly found by humans in their investigations. Intelligent systems responsible for the generation of collective intelligence are called Collective Knowledge Systems (CKSs). In this type of systems the users are producers and consumers of knowledge at the same time. In fact they enter the information and "eat" the knowledge generated by machines through a inference processes over data.

The diversity of operations and functionalities needs diversified data structures and aggregations. Data can be structured like database, semi-structured like XML documents or totally unstructured like free texts. A preliminary operation to perform is to process the unstructured users information. In order to achieve this goal a CKS need for implementing a strong data structuring. By combining unstructured and structured data it could provide a substrate for discovery of new knowledge that is not contained in any one source. In many cases this structure is introduced through the use of a relational database and/or XML-based technologies.

Another crucial aspect is the way people interact in the environment to communicate and to exchange information. Traditionally the field of research deputed to improve interaction between users and machine is the Human Computer Interaction (HCI). Human Computer Interaction's main purpose is the definition of paradigms and tools that enable humans to attain an optimal level of comfort while using computer based devices through proper interfaces. The new way people interact is mediated by machines and lead to a new research field called Human to Human Interaction (HHI). HHI is a challenging new domain where networked information systems and intelligent environments converge for the purpose of helps people to cooperate for any task, any time and any where. It describes how today human interaction is largely indirect and mediated by an increasingly wide range of technologies and devices (HHI can be seen as a human-to-computer-to-human interaction). All this peculiarities allows easier knowledge exchange between people

and machines, increasing the effectiveness of Collective Knowledge Systems.

These methodologies has been tested in a field of research developed in collaboration with "Dipartimento di Scienze Filologiche e Linguistiche" of the University of Palermo for the ALS (Atlante Linguistico Siciliano - Linguistic Atlas of Sicily) project. This is one of the first sociolinguistic projects across Italy and Europe. It was started in the early nineties from an intuition of professor Giovanni Ruffino (Ruffino and D'Agostino (2005)). Sociolinguistics studies the language as a dynamic and complex field changing over space and time. It incorporates knowledge and information from many other disciplines, such as anthropology, sociology, psychology and lately computer science and information processing.

The project is currently a joint effort with research units in Palermo, Catania and Messina leaded by the Dipartimento di Scienze Filologiche e Linguistiche of the University of Palermo. As stated, the project goal, from an operative point of view, is to exploit different type of phenomena related to phonetic, lexical, morpho-syntactic and textual aspects of language. The focus is on the relations between the evolution, over space and time, of the usage of regional Italian and Sicilian dialect. A relevant aspect is the dissemination of the produced results as a way to keep track of the social evolution through language. The dissemination is performed through specific publications and reports.

This on-going collaboration process is a perfect example of a domain hybridizing process, enabling the training on-the-field of a joint group of researchers who, coming from the peculiarly different scientific and cultural domains pertaining to the project, participate to the constitution of the core of a local humanities computing community. The project is extensively interdisciplinary and the progress, made from a technology point of view, are mostly in the direction of building an inter-operable infrastructure.

The ALS framework is the virtual linguistic laboratory, developed as a web application, built to support research's investigation e collaboration. The purpose of the framework is to model the entire process regarding the different steps of data acquisition, data transformation and research hypotheses verification in the ALS project.

The socio-linguistic researcher, that is the main actor of the ALS framework, has to acquire information in many formats: multimedia data, audio data, question-answer (textual) from a particular questionnaires (ALS questionnaires). The first step of the process is the acquisition of data, that is related to the particular media type. In the second phase data are enriched with a first level of information produced by researchers. A typical example is the linguistic annotation process, according to which a domain expert introduces information related to the particular aspect that he wants to deep in. The new information are stored to be retrieved in a second time, using the query tools that the ALS framework provides.

The developed framework integrate different technologies to solve the presented

issues and uses the web technologies as a container to keep track of the overall process for the researcher. In this way he can propose a multi-user system, that allow collaboration between linguistic researcher by sharing of results and methodologies of data aggregation. So the collaborative aspect of the platform is a milestone because it supports data analysis from different linguistic points of view.

One of the major issues to be taken into account is the level of usability of the system. An high degree of usability ensures a simpler learning curve for users that are mainly skilled in linguistic and are not used to perform complex statistical experiments. HHI methodologies help us to obtain good level of usability and a enjoyable user experience.

The ALS project represents also an excellent testing ground for definition of new intelligent methodologies for knowledge discovery in database (KDD). KDD is a process whose goal is the discovery of new patterns derived from semi-automatic processing of data in one or more databases. These patterns are usually searched in order to give support to the business decision-making. The most important step of the KDD process is called Data-Mining which provides for the implementation of specific algorithms to extract new information from the system knowledge base.

The ALS framework is able to connect high-level concepts, defined in the ontological level, with instances of structured data, which organize the results of the ALS questionnaires. In this perspective, the system is able to map the characteristics of the concepts by means of a semi-structured representation of the basic elements based on XML. The user can combine high-level concepts between them by logical relationship, so he can perform a query to retrieve information from the system. Moreover the researcher can save the query result as a new concept that successively can be used to make another investigation over data. So the knowledge is increased as a function of the users interaction, that are able to define, and subsequently share, new data patterns that increase the initial knowledge base.

1.1 Dissertation Outline

The rest of this thesis is organized as follows. In the next section there is the list of publications to which I gave my contribution during the Ph.D. study period. The chapter 2 is an overview about the scientific basis of this thesis. In the specific, the section 2.1 is a short dissertation about Knowledge Discovery in database, the section 2.2 argues Collective Intelligence principles and finally section 2.3 introduces the new research domain of Human-to-human interaction. In the chapter 3 there is a description of Linguistic Atlas of Sicily project and an overview of ALS framework, the HHI platform deployed to support linguistic research on own work. Then some conclusions and future works are presented in 4. While the appendix A describe a platforms that is a good example of HHI methodologies that involves

the use of touch screen devices of mobile devices for the information fruition in Large Social Events.

1.2 Publications

Journal papers:

- Gentile, A.; Andolina, S.; Massara, A.; Pirrone, D.; Russo, G.; Santangelo, A.; Sorce, S.; Trumello, E., "QRouteMe: A multichannel information system to ensure rich user-experiences in exhibits and museums". (2012) Journal of Telecommunications and Information Technology, 2012, vol. 1, pp. 58-66

Book chapters:

- Pirrone, D.; Russo, G.; Gentile, A.; Pirrone, R., "Collective Reasoning over Shared Concepts for the Linguistic Atlas of Sicily". (2013) Inter-Cooperative Collective Intelligence: Techniques and Applications. ISBN: 978-3-642-35015-3, Series "Studies in Computational Intelligence", Eds. Springer, Berlin Heidelberg, vol. 495, pp 403-425

International Conference Proceedings:

- Pirrone, D.; Andolina, S.; Santangelo, A.; Gentile, A.; Takizawa, M., "Platforms for Human-human Interaction in Large Social Events", Broadband, Wireless Computing, Communication and Applications (BWCCA), 2012 Seventh International Conference, pp. 545-551, DOI: 10.1109/BWCCA.2012.96

- Andolina, S.; Pirrone, D.; Russo, G.; Sorce, S.; Gentile, A., "Exploitation of Mobile Access to Context-Based Information in Cultural Heritage Fruition". (2012) Broadband, Wireless Computing, Communication and Applications (BWCCA), 2012 Seventh International Conference, pp. 322-328, DOI: 10.1109/BWCCA.2012.60. BEST PAPER AWARD

- Gentile, A.; Andolina, S.; Massara, A.; Pirrone, D.; Russo, G.; Santangelo, A.; Trumello, E.; Sorce, S., "A Multichannel Information System to Build and Deliver Rich User-Experiences in Exhibits and Museums," Broadband and Wireless Computing, Communication and Applications (BWCCA), 2011 International Conference, pp.57-64, DOI: 10.1109/BWCCA.2011.14

- Pirrone, D.; Russo, G.; Gentile, A.; Pirrone, R.; Gaglio, S., "The ALSWEB Framework: A Web-based Framework for the Linguistic Atlas of Sicily Project," Advanced Information Networking and Applications (AINA), 2011 IEEE International Conference, pp. 571-577, DOI: 10.1109/AINA.2011.77

Chapter 2

Theoretical background

The current chapter illustrates the research areas that are the theoretical basis of this dissertation. At first a study of the state of the art in the field of knowledge discovery in database is presented, a topic which is the main scientific foundation of this studies, then Collective Intelligence and Human-to-Human Interaction are presented. The use of management techniques borrowed from Collective Intelligence and a particular attention to the theme of Human-to-Human Interaction in a social web system has allowed to explore new methodologies to build emerging knowledge that greatly contributes increasing the knowledge base of an human-machines system. The fields have been investigated in relation to find new ways to add knowledge to complex systems via human interactions.

2.1 Knowledge Discovery in Database

The majority of companies operating in different sectors must often cope with the growing amount of data that is accumulated over time and that is stored in special systems such as databases. These data represent a valuable resource for the discovery of useful knowledge and to provide decision support in business processes. One example of a supermarket that, through the use of a bar code reader, is capable to storing in a database the entire list of customers purchases. The supermarket might get some interesting information regarding products sold through an analysis of the data which is in their possession in order to improve businesses.

The conventional method of acquiring knowledge from data is based on the data analysis and their interpretation. For example, in the health sector the specialists analyze the current trends and changes in the data on a quarterly basis, preparing a report for the healthcare organization. This report will be the starting point on the basis of which to start the decision-making and planning of health

care management. Similarly planetary geologists analyze images of planets and asteroids captured by the sensors having the need to locate and accurately categorize the various geological phenomena of interest, such as impact craters.

Knowledge Discovery in Databases was used at the first KDD workshop in 1989 (Piatetsky-Shapiro (1991)) to emphasize that knowledge is the end product of a data-driven discovery. It consists in development of methods and techniques for making sense of data. It has evolved from the intersection of research fields such as: Machine Learning, Pattern Recognition, Databases, Statistics, Artificial Intelligence, Knowledge Acquisition for expert systems, Data Visualization, High-Performance Computing.

The term Knowledge Discovery is defined as the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data (Fayyad et al. (1996)). The data are a set of facts, such as tuples in a relational database, while the pattern is an expression in a given language, which describes a significant subset of data. Hence, extracting a pattern also designates fitting a model to data or finding structure from data; or, in general, making any high-level description of a set of data. The nontrivial attribute implies that some search or inference is involved. The discovered patterns should be valid on new data with some degree of certainty. The patterns have to be novel (at least to the system and preferably to the user) and potentially useful, that is, lead to some benefit to the user or task. Finally, the patterns should be understandable, if not immediately then after some post-processing.

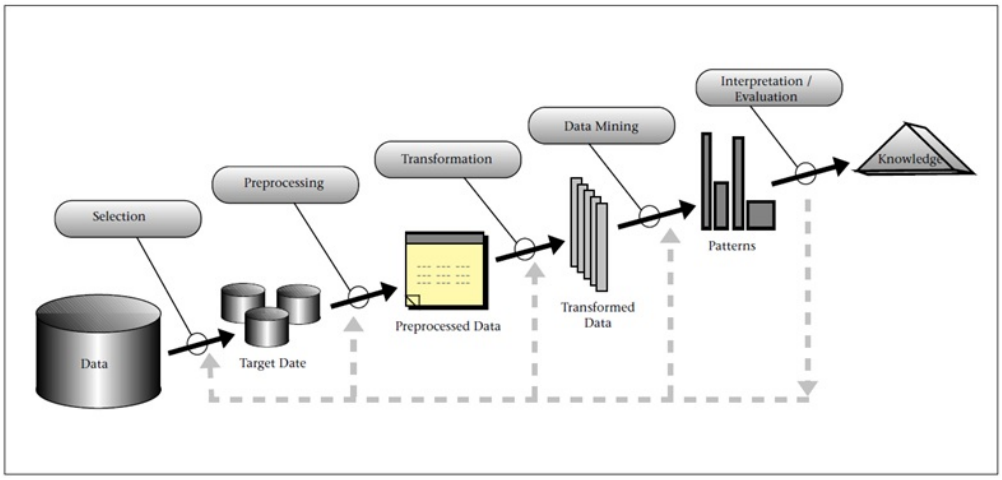

Figure 2.1: Knowledge Discovery phases

KDD is divided into several steps involving the preparation of the data, searching for patterns, analysis and refinement of knowledge (figure 2.1). Data Mining is the key phase that consists of applying data analysis and discovery algorithms

that, under acceptable computational efficiency limitations, produce a particular enumeration of patterns. It is important to say that the space of patterns is often infinite, and the enumeration of patterns involves some form of search in this space. Practical computational constraints place severe limits on the subspace that can be explored by a data-mining algorithm.

The initial step of the KDD process consists in the development of application domain understanding, relevant priori knowledge, and end user objectives. Then it will be possible carry out the second phase, the selection of a subset of data based on which it will search for knowledge. The third phase is characterized by data cleaning and data preprocessing: basic operations such as removing noise or outliers if appropriate, collecting the necessary information to build a noise model, development of strategies to manage missing data and to manage time-varying data. The step of data reduction and projection is aimed at the representation of the data in an appropriate way, in relation to objectives of the research, while size reduction and the use of processing methods has the objective of reducing the actual number of variables to be the research process.

The choice of the data mining task (classification, regression or clustering) is a critical step in the KDD process that precedes the selection of the Data Mining algorithms and selection methods that make the search for patterns in the data. This phase includes the decision about which models and parameters may be appropriate and matching a particular data mining method with the overall criteria of the KDD process (The end user might be more interested in the understanding of the model rather than its predictive capabilities)

2.1.1 The Data-Mining Step

The data-mining component of the KDD process is considered the most important phase of the entire process. It often involves repeated iterative application of specific data-mining methods. It is primary to detect the knowledge discovery goals that are defined by the intended use of the system. it is possible to distinguish two main types of goals: verification and discovery. With verification, the aim of system is verifying the user's hypothesis, while with discovery, the system autonomously finds new patterns. The discovery can be further divided into prediction, where the system finds patterns for predicting the future configuration of some entities, and description, where the system finds patterns for better explain the selected data meaning to a user in a human-understandable form. Although the boundaries between prediction and description are not sharp (some of the predictive models can be used to descriptive methods and vice versa), the distinction is useful for understanding the overall discovery goal.

Data mining involves fitting models to observed data that play the role of inferred knowledge. This models reflect useful or interesting knowledge as a part

of the overall, interactive KDD process where subjective human participation is often required. Two primary mathematical formalisms are used in model fitting: statistical and logical. The statistical approach allows for non-deterministic effects in the model, whereas a logical model is purely deterministic. The statistical approach to data mining tends to be the most widely used basis for practical data-mining applications given the typical presence of uncertainty in real-world data processes.

Most data-mining methods are based on tried and tested techniques from machine learning, pattern recognition, and statistics like: classification, clustering, regression, and so on. The different algorithms set can often be bewildering to both the novice and the experienced data analyst. Many data-mining methods advertised in the literature, there are really only a few fundamental techniques. The underlying model representation being used by a particular method typically comes from a composition of a small number of well-known options: polynomials, splines, kernel and basis functions, threshold-Boolean functions, and so on. Thus, algorithms tend to differ primarily in the goodness-of-fit criterion used to evaluate model fit or in the search method used to find a good fit. In fact most methods can be viewed as extensions or hybrids of a few basic techniques and principles.

The primary methods of data mining methods can be viewed as consisting of three primary algorithmic components:

1. Model representation
2. Model evaluation
3. Search method

Model representation is the language used to describe discoverable patterns; if the representation is too limited, then an accurate model for the data can't be produces. It is important to comprehend that specific model representations are being made by a particular algorithm and it's unlikely good to another algorithm. Increased representational power for models increases the danger of over fitting the training data, resulting in reduced prediction accuracy on unseen data.

Model evaluation criteria consists in a quantitative statements or fit functions that make an estimate of how well a particular pattern meets the goals of the KDD process. For predictive models the evaluation is often made by a empirical measure of accuracy on some test set. Descriptive models can be evaluated along the dimensions of predictive accuracy, novelty, utility, and understandability of the fitted model.

Search method consists of parameter search and model search. Once the model representation and the model evaluation criteria are fixed, then the data-mining algorithm is became an optimization problem. In parameter search, the

algorithm must search for the parameters that optimize the model evaluation function. Usually the model representation is changed so that a family of models is considered.

2.1.2 Data-Mining Methods

In literature many method to implement data mining algorithm exist. To better understand how this methods work, it is used a simple example to make some of the notions more concrete. Figure 2.2 shows a two-dimensional data set consisting of 23 cases. Each point on the graph represents a person who has been given a loan by a particular bank in the past. The horizontal axis represents the income of the person while the vertical axis represents the total personal debt. The data have been classified into two classes: the 'X' point represent persons who have defaulted on their loans and the 'O' represent persons whose loans are in good status with the bank. This simple data set certainly contains useful knowledge from the bank point of view, that can be discovered.

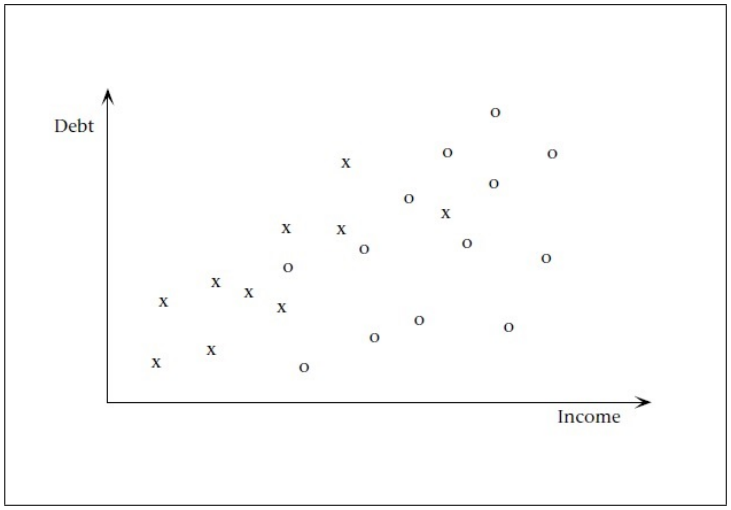

Figure 2.2: Example of two dimensional data set

A possible method to implement data mining is **Classification** that consists in learning a function that maps (classifies) a data item into one of several predefined classes (Weiss and Kulikowski (1990)). In knowledge discovery applications examples of classification methods used as part of include the classifying of trends in financial markets (Apte and Hong (1995)). Figure 2.3 shows a simple partitioning of the loan data into two class regions; note that it is not possible to separate the classes perfectly using a linear decision boundary. The bank might want to use the classification regions to automatically decide whether future loan applicants will be given a loan or not.

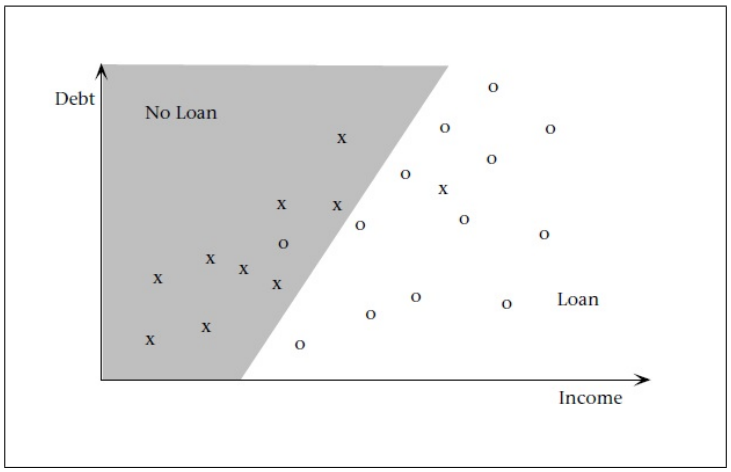

Figure 2.3: simple partitioning of data using a linear decision boundary

Regression method consists in learning a function that maps a data item to a real-valued prediction variable. Figure 2.4 shows the result of simple linear regression where total debt is fitted as a linear function of income: the fit is poor because only a weak correlation exists between the two variables.

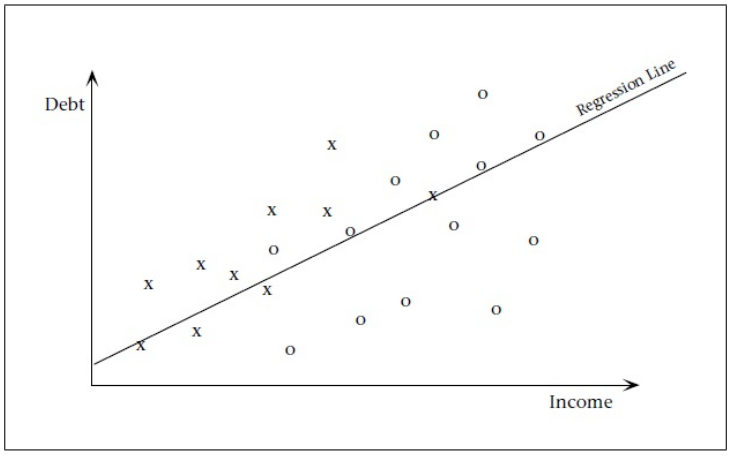

Figure 2.4: result of simple linear regression

Clustering is a common descriptive task where one seeks to identify a finite set of categories or clusters to describe the data (Titterington et al. (1985)). The categories can be mutually exclusive or consist of a richer representation, such as hierarchical or overlapping categories. Figure 2.5 shows an example of clustering method applied to the loan data set. Note that the tree identified clusters overlap, allowing data points to belong to more than one cluster. All points of graph have been represented by a '+' symbol to indicate that the class membership is no

known. Closely related to clustering is the task of probability density estimation, which consists of techniques for estimating from data the joint multivariate probability density function of all the variables or fields in the database (Silverman (1986)).

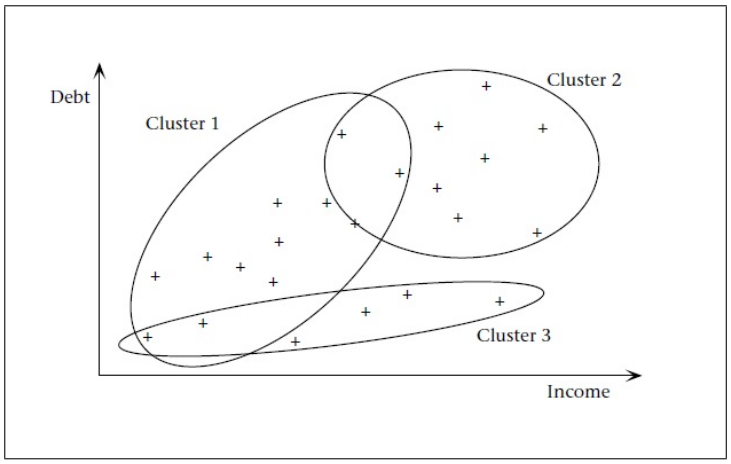

Figure 2.5: example of clustering method applied to the loan data set

Summarization involves methods for finding a compact description for a subset of data. A simple example would be tabulating the mean and standard deviations for all fields. More sophisticated methods involve the derivation of summary rules, multivariate visualization techniques, and the discovery of functional relationships between variables. This method are often applied to interactive exploratory data analysis and automated report generation.

Dependency modeling consists of finding a model that describes significant dependencies between variables. Dependency models works at structural and quantitative level. The structural level of the model specifies, often in graphic form, which variables are locally dependent on each other, while the quantitative one specifies the strengths of the dependencies using some numeric scale. For example, probabilistic dependency networks use conditional independence to specify the structural aspect of the model and probabilities or correlations to specify the strengths of the dependencies (Glymour (1987)), (Heckerman, 1997).

Change and deviation detection focuses on discovering the most significant changes in the data from previously measured or normative values ((Berndt and Clifford, 1996), (Guyon et al., 1996), (Matheus et al.) and (Basseville and Nikiforov, 1993)).

2.2 Collective Intelligence

The Collective intelligence idea was born in 1963 such as the goal of visionaries throughout the history of the Internet. Douglas Engelbart, who invented groupware, mouse, and form of hypertext designed for collective knowledge, wrote: "The grand challenge is to boost the collective IQ of organizations and of society" (Engelbart (1963)). His Bootstrap Principle was about a human-machine system for simultaneously harvesting the collected knowledge for learning and evolving our technology for collective learning. Collective intelligence can be taken as a scientific and societal goal, and throw Internet the possibilities, that it happens now, are very high. The key, as the visionaries have seen, is a synergy between human and machines. In this type of system, both human and machine contribute actively to the process of building intelligence. People are the producers and customers: they are the source of knowledge, and they have real world problems and interests. Machines are the enablers: they store and remember data, search and combine data, and draw mathematical and logical inferences. People learn by communicating with each other, and often create new knowledge in the context of conversation. Internet makes it possible for machines to help people create more knowledge and learn from each other more effectively.

Collective Intelligence (CI) can be defined as the surplus of intelligence created by a group or a community emerging by the contributors' interactions using a system and ultimately the Internet . There are other way to define CI like crowd sourcing (Howe (2006)), wisdom of crowd (Surowiecki (2004)) and smart mobs (Rheingold (2003)). All this are united by the fact that the final output is achieved the collective efforts of individual or groups of them. The MIT Center for Collective Intelligence conducts research on how new communications technologies are changing the way people work together. It has this basic research question: "*How can people and computers be connected so that-collectively-they act more intelligently than any individuals, groups, or computers have ever done before?*" (MIT).

To produce collective intelligence a key component is a *Collective Knowledge System* (CKS) as a system where "machines enable the collection and harvesting of large amounts of human-generated knowledge" (Gruber (2008)). A CKS incorporates some important elements that are intelligent users able to produce knowledge, an environment allowing exchange of relevant information and social interactions between users and a service of knowledge retrieval such as a search engine to find relevant information. One of the most important features of such service is the capability to produce emergent knowledge which is the knowledge that is not directly found by humans in their investigations.

The definition of such systems involves the ability to keep track of interactions between users in the network which is known as social network analysis.

This lead to a new way to consider networks of people producing, modifying and exchanging information. The digital environment become in such way a Digital Ecosystem (Di Maio (2008)) meaning that like natural ecosystem rules and are affected and affects other components. Numerous attempts to produce and realize systems according to the definition of collective intelligence has been carried on in recent years. The main applications involve definition of social networks explicitly modelled according to the Collective Intelligence principles (Lek et al. (2009)) or the application of collective intelligent principles to build systems able to perform specific tasks (Lertnattee et al. (2009), Mizuyama and Maeda (2010) and Maher and Kourik (2008)).

With the high diffusion of Social Web among people, there are millions of humans that offer their knowledge to human computer systems. It means that the information about this people is stored, searchable, and easily shared. The goal for the next generation of the Social Webs is to define new methodology to enhance useful reasoning with the large amount of data that is put online by systems' users. True collective intelligence can emerge if the data collected from all those people is aggregated and recombined to create new knowledge and new ways of learning that individual humans cannot do by themselves.

2.2.1 Collective knowledge systems

It is possible to define collective knowledge systems as: human to human systems (see section 2.3) in which machines enable the collection and harvesting of large amounts of human-generated knowledge. This type of system are characterized by the following features:

1. **User-generated content**. The knowledge base of the system is provided by humans participating in a social process with the purpose of sharing knowledge.

2. **Human-machine synergy.** The combination of human and machine provides a capacity to provide useful information that could not be obtained otherwise. People are the producers and customers: they are the source of knowledge, and they have real world problems and interests. Machines are the enablers: they store and remember data, search and combine data, and draw mathematical and logical inferences.

3. **Increasing returns with scale.** As more people contribute, the system gets more useful. The system of rewards that attracts contributors and the computation over their contributions is stable as the volume increases.

4. **Emergent knowledge.** The system enables computation and inference over the collected information, leading to answers, discoveries, or other results

that are not found in the human contributions. Emergent knowledge is the product of a technology for assembling large sources of human-generated content in such a way that computations can discover and conclude new useful information.

2.2.2 Need for structured data

The Emergent knowledge is the peculiar difference of collective knowledge system from the most common collected knowledge system. To move closer to Emergent Knowledge, it is necessary to enrich user-contributed data with structured data. Semantic Web technologies can add structured data related to the content of the user contributions. By combining structured and unstructured data it could provide a substrate for the discovery of new knowledge that is not contained in any one source. The standards and infrastructure of the Semantic Web can enable data sharing and computation across independent, heterogeneous Social Web applications (see section 2.3.1).

The essential difference between the classic Web and the Semantic Web is that structured data is exposed in a structured way. There are three basic way to do this:

- expose data that is already in the databases used to generate HTML pages
- extract the data retrospectively from user contributions
- capture the data as people share their information

The first approach consists in exposing the structured data, stored into relational database, that are used to build dynamically the unstructured web pages. A possible technique is make accessible the structured data in web pages using standard formats. This requires that the web developer provides to add this information to web pages during the web-site development. For example, Social Web sites could expose their links to users as FOAF data, which is a Semantic Web convention for representing personal contact information (Brickley and Miller (2007)).

To extract structured data from unstructured user contributions there are several techniques (Auer and Lehmann (2007), Mooney and Bunescu (2005) and Suchanek et al. (2007)). For example, elaborating unstructured user contributions, it is possible to identify people, companies, and other entities with proper names, products, possible relations the users are interested in (Agichtein and Gravano (2000) and Cafarella et al. (2005)), or instances of questions being asked (Lita and Carbonell (2004)).

There are also semi-automatic techniques for identifying classes and relations from unstructured user contributions, although these are a bit noisier than the directed pattern matching algorithms (Suchanek et al. (2006), Lin and Pantel (2002), Pantel and Ravichandran (2004), Pantel (2006) and Tokunaga et al. (2005)). What is interesting is that these techniques can be used to enrich the unstructured user data with structured data, that represents some of the entities and relationships mentioned by users. For example if a Wikipedia web page has a reference to book using its ISBN number, it could link to the book in structured databases of books and be used to call APIs for obtaining it (Bizer et al. (2007)). More sophisticated examples for extracting references to named entities and factual assertions can also be applied. It is important to note that all these techniques require open data access and APIs to have a real impact on the Social Web.

The third approach is to obtain structured data directly from users on the way. In practice this technique consist in to give users tools to add structured fields and make class hierarchies over the unstructured data. This is inappropriate for Social Web application, because the users in this space are not there to create databases but to socialize and be fun. However, there is a way that makes this practice useful for the users. It is a method that reward the users for the effort generated by this thankless task, that gives in return to them a social value for the task to add structure to contents. We call this technique snap-to-grid.

The term "snap-to-grid" refers to an interaction pattern that many of us use without consciously thinking about it. When you draw a shape on the screen in a drawing program, by default the system finds the nearest point in a discrete grid of locations to align your edges. Similarly, when you type text by a word processor program, the system is automatically snapping your text to the nearest word in the "grid" provided by the dictionary. Email programs do a similar "completion" on addresses typed into the address field. For the user snap-to-grid creates a more attractive and useful product. For example for the email application, the service on addresses typed is indispensable to achieve all of the messages sent by the same person, or at least from the same address. The "snap-to-grid" procedure result is a mix of structured and unstructured data, which has far more value when aggregated into collections.

Snap-to-grid assumes that there is a structure to data (such as constraints on its form or values), and helps users enter data within that structure. It is important to combine a snap-to-grid interface for soliciting structured data with motivations for providing this data. For example, there are tools for adding structure toWikipedia (Völkel et al. (2006)), but they depend on voluntary compliance. An interesting approach is to combine data entry with a social system that structures the behavior. For example in Ahn (2006) a intelligent and useful (for adding structure to data) game is described. In this games people are rewarded for teaching the computer things such as what to label an image. In practice the aim of the game

is the following: an entire image, or a well-defined region of the image, must be label whit a word by the players that try to match the label of other players. The motivational structure of the game and the large number of players leads to quality of content.

2.3 Human-to-human interaction

Human-to-human interaction (HHI) is a challenging new domain where networked information systems and intelligent environments surrounding people converge for the purpose of better satisfaction of users' requirements and anticipation of their needs (Gentile et al. (2011)). HHI is derived from human-computer interaction (HCI). Histirically, the main goal of researchers HCI field was to set up suitable I/O interfaces for personal computers. Such interfaces had to be as intuitive to use and as easy to learn as possible for human beings. Results achieved in this field have resulted in hardware and software objects that are now part of our everyday life, such as mices, graphic tablets, graphical users interfaces. Widespread distribution of PCs and their shared acceptance in everyday life for most of us is its clearest demonstration. Nowadays, the growing and growing diffusion of smart personal mobile devices makes the internet more pervasive than ever, accessible through devices in anyone's pocket. People are able to interact whenever they mutually agree, even when separated by long distances. This means that humans end up interacting more and more with other humans through a computing device, as distances grow longer and communication needs become more and more pressing. In other words, a growing portion of HHI can be seen as a human-to-(computer-to-)human interaction.

As shown in Figure 2.6, HHI can be enabled by means of direct (in-person) human interactions in real environments as well as by means of virtual interacting representations. In the first case, HHI interfaces represent devices and actuators for natural interactions, involving the basic human senses. In the latter case, HHI interfaces can be seen as input devices to generate a virtual representation of the involved subject. The path (human - input technology - virtual representation - output technology - human) is the typical path of a social network-based interaction. Several human-to-computer interfaces can be used and combined to address the interaction issues, according to the environmental conditions and the user preferences.

2.3.1 Social network

A social network is a social structure made up of configuration of people, which are connected to one another through one or more specific types of interdependency,

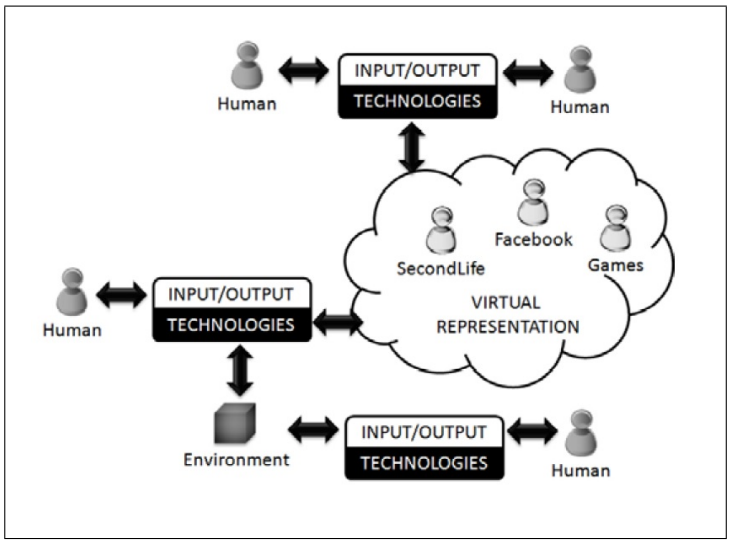

Figure 2.6: HHI throw virtual representations

such as friendship, financial exchange, family relationships, or relationships of beliefs, common interest, knowledge, etc. Online Social Networking (OSN) sites are virtual places that implements social networks in which people gather to communicate, share and discuss ideas. The social network system is represented in Figure 2.6 with the virtual representation cloud, in which virtual profile of the human are stored.

The goal is to provide user with best formed computer interfaces to satisfy needs of users. Some well-known OSNs, networks foster relationship building and communications among those involved. The social software definition encompasses a set of software systems that allow users to interact and share data.

This computer-mediated communication has become very widespread with social sites like MySpace, Facebook and Twitter, media sites like Flickr and YouTube as well as commercial sites like Amazon.com and eBay. Social software are also described with the terms Web 2.0 (DiNucci (1999) and Murugesan (2007)), which define the set of applications that facilitate interactive information sharing, interoperability, user-centred design, and collaboration on the web. Much of what is present in literature about Web 2.0 relates to its use by individuals because of the speed of development in web technologies. In fact, Web 2.0 has become an important internet application with the advent of the integration of social interaction and web technologies.

On the internet conversations between hundreds or thousands of users have become a usual practice. People exchange thousands of messages in online conversation, this type of interaction that is large scale, public, many-to-many and persistent is named very large-scale conversation (VLSC) (Sack (2000)). The size of

VLSCs make problematic for participants to understand and manage interaction. Sack (2000) discussed the design criteria to transform social scientific representations into interface devices. The issue is illustrated with the description of the conversation map system, which is a system for browsing VLSCs.

The best example of global social networking community is Facebook, which is a virtual community opened in February 2004. As of October 2013 Facebook has more than 1 milliard active users in a month (58% are connected every day, the others enter about one time for month).

Chapter 3

Linguistic Atlas of Sicily

In this chapter, collective intelligence principles are applied in the context of the Linguistic Atlas of Sicily, an interdisciplinary research focusing on the study of the Italian language as it is spoken in Sicily, and its correlation with the Sicilian dialect and other regional varieties spoken in Sicily. The project has been developed over the past two decades and includes a complex information system supporting linguistic research; recently it has grown to allow research scientists to cooperate in an integrated environment to produce significant scientific advances in the field of ethnologic and sociolinguistic research. The definition of an integrated methodology able to perform a comprehensive analysis in the different fields of sociolinguistic is the ultimate goal of the Linguistic Atlas of Sicily project.

This is one of the first sociolinguistic projects across Italy and Europe. It was started in the early nineties from an intuition of professor Giovanni Ruffino (Ruffino and D'Agostino (2005)). Sociolinguistics is the the study and the explanation of possible connections between society behaviors and language evidences. Sociolinguistics study the language as a dynamic and complex field changing over space and time. Sociolinguistics incorporate knowledge and information from many other disciplines, such as anthropology, sociology, psychology and lately computer science and information processing.

The project is currently a joint effort with research units in Palermo, Catania and Messina leaded by the Dipartimento di Scienze Filologiche e Linguistiche of the University of Palermo. As stated, the project goal from an operative point of view is to exploit different type of phenomena related to phonetic, lexical, morpho-syntactic and textual aspects of language. The focus is on the relations between the evolution over space and time of the usage of regional Italian and Sicilian dialect. A relevant aspect is the dissemination of the produced results as a way to keep track of the social evolution through language. The dissemination is performed through specific publications and reports. The on-going collaboration process is a perfect example a domain hybridizing process, enabling the training on-the-field

of a joint group of researchers who, coming from the peculiarly different scientific and cultural domains pertaining to the project, participate to the constitution of the core of a local humanities computing community.

The project is extensively interdisciplinary and the progress made from a technology point of view are mostly in the direction of building an inter-operable infrastructure. The infrastructure is organized to allow exchange of information and knowledge between users providing tools and methodologies to allow collective reasoning over shared concepts. The project uses different types of data (structured, unstructured, multimedia) so a first aspect is related to data integration and interoperability. Another important aspect is the exchange of information resulting in the process of data aggregation. Information visualization is the last step of the overall process.

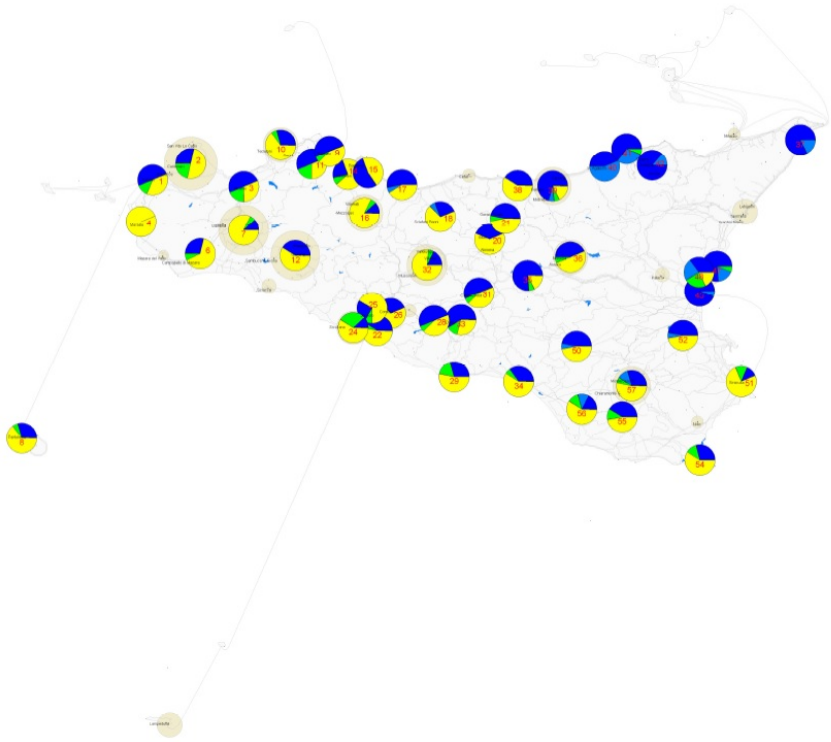

Figure 3.1: Statistical map.

The sociolinguistics field has been investigated over the years in different nations especially where the preservation of a national identity is an urgency. Numerous projects have been realized and some are still working. Usually the dimension of the project is related to the purposes of the project itself. A common

goal is to produce atlas representing particular linguistic occurrences. A good example is the Linguistic Atlas of Dolomitic Ladinian and neighbouring Dialects (https://www.sbg.ac.at/rom/people/proj/ald/aldhome.htm). This project is related to preservation and study of ladinian dialect which is an ancient language spoken in some european regions. Another relevant example is the Linguistic (and ethnographic) Atlas of Castille - La Mancha (http://www.linguas.net/). In this atlas of the central region of the Iberian Peninsula, phonetic, lexical and grammatical information is offered. The data has been collected in situ, over the course of several years, through interviews carried out by questionnaires and through recorded conversations.

A transnational project is the PRESEEA project. PRESEEA is the “Project for the Sociolinguistic Study of Spanish from Spain and America". The goal is to coordinate sociolinguistic researchers from Spain and the Hispanic America in order to make possible comparisons between different studies and materials, as well as a basic information exchange. The main aim is to create a spoken language Corpus with sufficient guarantees and rich in terms of linguistic information. A more complete project has been carried out over the year from the Copenaghen University. The project is called Lanchart (http://lanchart.hum.ku.dk/aboutlanchart/) and is working for over three decades with the definition of a multidisciplinary centre to keep track of linguistic evolution in danish language.

3.1 Information Organization in the ALS framework

The ALS (Linguistic Sicilian Atlas) framework has been developed as a virtual laboratory and deployed as a web application. Each component of the framework is deputed to the definition of a set of functionalities that support the researchers in their jobs. In figure 3.2 the main flow of the process is depicted in order to define the pipeline to acquire, transform and use data related to the project.

3.1.1 Data Integration

Data integration is a preliminary task to perform to allow information exchange and knowledge discovery. A common way to allow data integration is to use a neutral format. To this purpose a wide range of XML-related technologies has been investigated. XML (Bray et al. (2006)) is becoming de-facto the universal language for data exchange over the web. Some standard interfaces like SAX (sax) and DOM (Wood et al. (2000)) have been defined to directly access the content of an XML document. The XML structure naturally discriminates between data and data structures. At present XML has a series of related technologies and

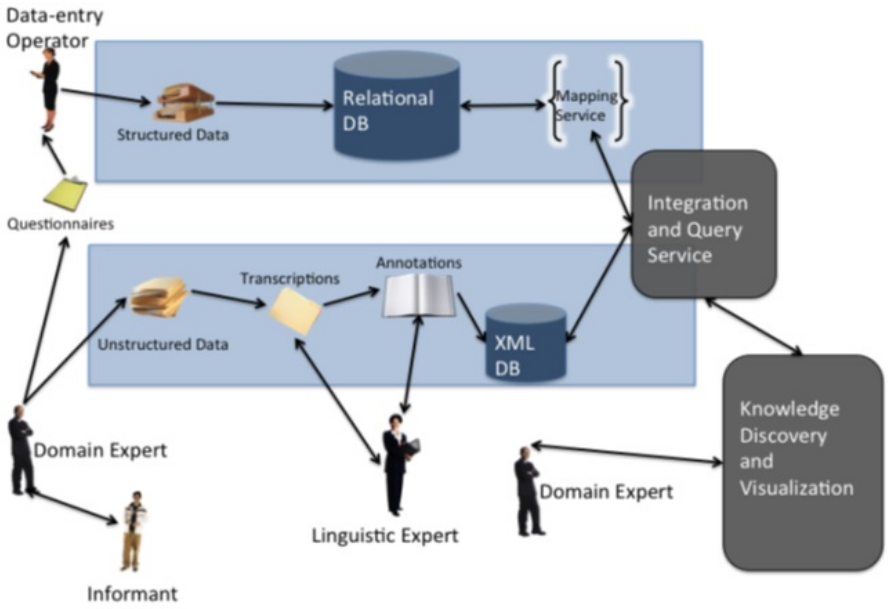

Figure 3.2: The main interactions of the overall project and the main interactions between users and system components.

formalisms useful to perform particular operations. Some examples are the DTD (Bosak et al. (2000)), and the XML Schema (Fallside and Walmsley (2004)), used to check and to validate the format of a particular XML document. To obtain a XML document from another one XSLT (Kay (2007)) transformations are used. XPath (DeRose and Clark (1999)) and XQuery (Siméon et al. (2007)) are used to query XML databases. The family of related technologies is growing and is able to perform more and more complex operations. The utilization of XML to perform annotation process starts from the past decade (Carletta et al. (2002)). The widely accepted approach is intended to define an abstract level of annotation that is independent from the physical organization of the process. The major problem is the multiple overlapping hierarchies, a peculiar characteristic of the nature of linguistic information. In the annotation process a single element can be analyzed in many ways and this leads to overlapping layers of annotations. An approach able to deal with the multidimensional nature in the annotation process is required. Many systems to perform annotation like GATE (Cunningham et al. (1996)), ATLAS (Bird et al. (2000)), or AWA (Artola et al. (2009)) have been developed. To solve the problem of the standoff annotation a system able to switch between the different annotation schemas is required. The other peculiar problem of the presented framework is related to the integration of different sources. An

XML based approach (Feng et al. (2008) and Draper et al. (2001)) to integrate different data sources is to prefer where some data are in an XML format or can be easily produced in an equivalent way.

3.1.2 Data Acquisition and Organization

Usually, the first step is data acquisition. Data have been acquired over the years trough a set of digitally recorded interviews that are derived from a multi-part questionnaire. The usual methodology developed in the ALS project allows to compare different results over time. The interviewed population is organized for families: a fixed number (usually five) of related people is chosen from different generations (usually three). The same number of families for a particular geographic point in Sicily is selected in order to have compatible clusters. In a standard questionnaire there are three main parts: a biographic section, a meta-linguistic section, and a linguistic section. The first section is composed of questions focused about personal information, statistics, level of instruction (personal and familiar), cultural consumes and other similar information useful to have a social characterization. The metalinguistic section is composed of questions investigating the perception and the self-perception about the two different linguistic codes (Italian language and Sicilian dialect) and the ability to move within the two codes. The third section regards specifically language and translation skills between the two codes or other related exercises.

After data acquisition, the information from the first and second section is stored into a relational database, while the third section is transcribed and than annotated directly by domain experts. In practice, a set of attributes is related to the transcript, to mark the linguistic phenomena. The result is a valid XML document responding to a XML schema, reflecting the level of annotation. There are two main components in a schema: a common part used to apply general information that is relevant for the retrieval part of the process, and a specific part that is defined by the domain expert. Different annotation schemas have been realized. Schemas for phonetic annotation work at a single word level, lexical annotation works at a phrase level and textual annotation works on a bigger portion of text. The last one makes the expert able to keep track of the interaction progress as well as people behaviors during the interview. Future definitions of new schemes are not precluded. New schemes can allow researchers to perform text analysis at a different level of abstraction, still not defined. The XML documents resulting from annotation, are stored in an XML-based DBMS.

The last part of the process is the retrieval of useful information. This process is defined in two main steps:

1. defining an integration process in conjunction with a query service, which

is able to retrieve information both from structured and semi-structured sources

2. knowledge discovery and visualization service, which is able to return useful information for users, to aggregate in a convenient way and to show information on a map for displaying geographic correlation between information items.

The entire framework is arranged into different sub-systems (Fig. 3.3). The main components are: the ALSDB (ALS Data Base) that manages relational data, the ALSML (ALS Meta Language) that manages semi-structured data, and the ALSGIS (ALS Geographic Information Systems) for visualization. More external components have been realized to allow data entry or annotation, and are complementary to the ALS framework.

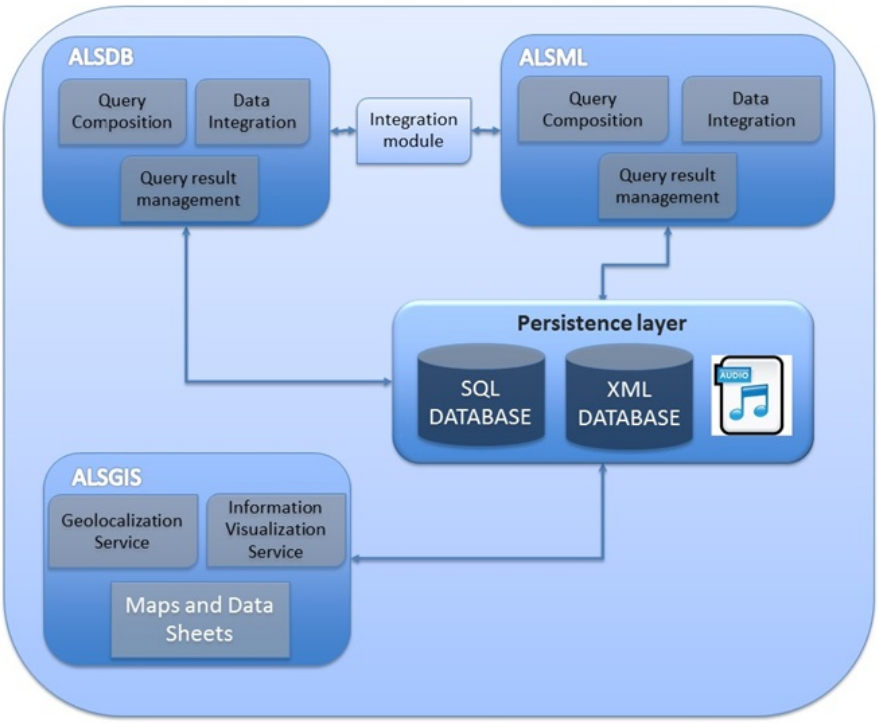

Figure 3.3: The three main components of the ALS project, the complementary tools and the data sources of the framework.

3.1.3 Information Definition Process

The process is an iterative one, and it allows cross validation between different users. An important feature is the possibility to share and prove hypothesis based

on empiric evidences found in the interrogation phase. The organization of the framework allows this exchange between different researchers.

The ALSDB is a central part in the process of knowledge discovery. A full automatic procedure able to retrieve the information, according to the user needs, has been realized (Russo et al. (2009)). The core process is the definition of a correspondence between the concepts in the ontology representing high-level hypothesis in the various domains of interest and the different views of a relational database. The user can be interested in different concepts with a different abstraction degree. He can reason with concepts that are at a very high level of abstraction (i.e. the perception of the user, his self-perception) or with basics concepts (i.e. the speaker's age, his degree of study). For this reason, the level of granularity of data doesn't allow a simple correlation between the main investigation variables, and even a simple query requires many lines of SQL code.

A possible solution to this problem is to use a correspondence XML file. The XML correspondences and the data of the DBMS are the inputs of the system to retrieve information. A correspondence between a concept and the database is defined through a mapping process, whose results are writ-ten in a XML file and validated according an ad-hoc built DTD. Every concept can be related to one or more table in the DBMS. For every table is possible to define the list of attributes that are involved in the concept definition. The correspondence is set also at a table level: the tables with attributes can be related with Boolean relations that are NOT, OR, AND.

After the definition of the correspondence between a concept and some tables it is possible to define some constraints for the mapping. Constraints are defined at a class level in the DTD with the list of possible attributes. Without defining constraints the mapping results in a default selection query. Using this paradigm, the user can compose a more complex query in an intuitive way throw a simple web interface, without writing lines of SQL code. The web interface allows users to refine results and save all the needed information. The result of a query is a set of tuples matching the selection criteria. The results can be saved as a new concept and be used as a starting point for further investigations. In this way the system produces an incremental knowledge base that is controlled by the domain experts. This knowledge base is accessible to other researchers to be validate and used in his domain of interest.

The ALSDB component works on a persistence layer data, composed by a set of XML files. For each concept, one persistent XML file exists that contains all the entries matching with the criteria defined inside it. This has several advantages. From the software engineering point of view, if objects are persistent, they can be easily defined as entities in the project (the domain classes). Another advantage is a better portability of the application: the presence of an abstraction layer from data gives the possibility to change easily their access policies, the interfaces or

the data drivers. The described persistence level has been implemented using the iBatis persistence framework (Begin et al. (2007b)).

3.2 The Main Framework Components

3.2.1 The ALSDB Component

The web interface for query composition is shown in figure 3.4. In the top left column the user can select the possible answers to a single question, defined into the questionnaire. In the top right column the user can choose an item in a list of concepts. Many types of concepts are available: some of them have been saved by other users or they are related to towns, scholar levels, family typologies, and so on. The selected concepts are added to the query panel to perform actual investigations.

This GUI allows to define complex research criteria by mixing concepts through boolean relationships. The user selects some the criteria that are inserted in the container of the temporary criteria to be aggregated with other ones. When selecting more than one concept, the user can put them in logical relationship (AND, OR), and insert them in the bottom container of the composed criteria. Infinite expressions can be inserted in the criteria container, and the user can either select one criterion to perform the query using the “Run Query" button, or re-use it in conjunction with other criteria through the “Use criterion" button. In the last case, the selected expression will be reinserted in the top container of temporary criteria, and it will be mixed in AND-OR relationships with other concepts, to create complex expressions increasingly.

Then user requests will be translated into a SQL query. The SQL code will have a “WHERE" clause that can be composed by a complex expression, using parentheses to manage priorities. A priority matrix is defined purposely to handle the incremental mechanism of expressions composition. Such a matrix contains all the information about concepts that are used in the query, the relationships between them, along with their priorities defined using parentheses. This matrix is created incrementally while the user combines the concepts using the GUI. The key point of the query generation service is precisely the priority matrix whose structure and contents are needed to generate the XML correspondence file and to build dynamic query.

3.2.2 The Transcription and Annotation Tools

The annotation tools are used to transform portions of the interview (in third part of questionnaires) into semi-structured documents, which contain information (an-

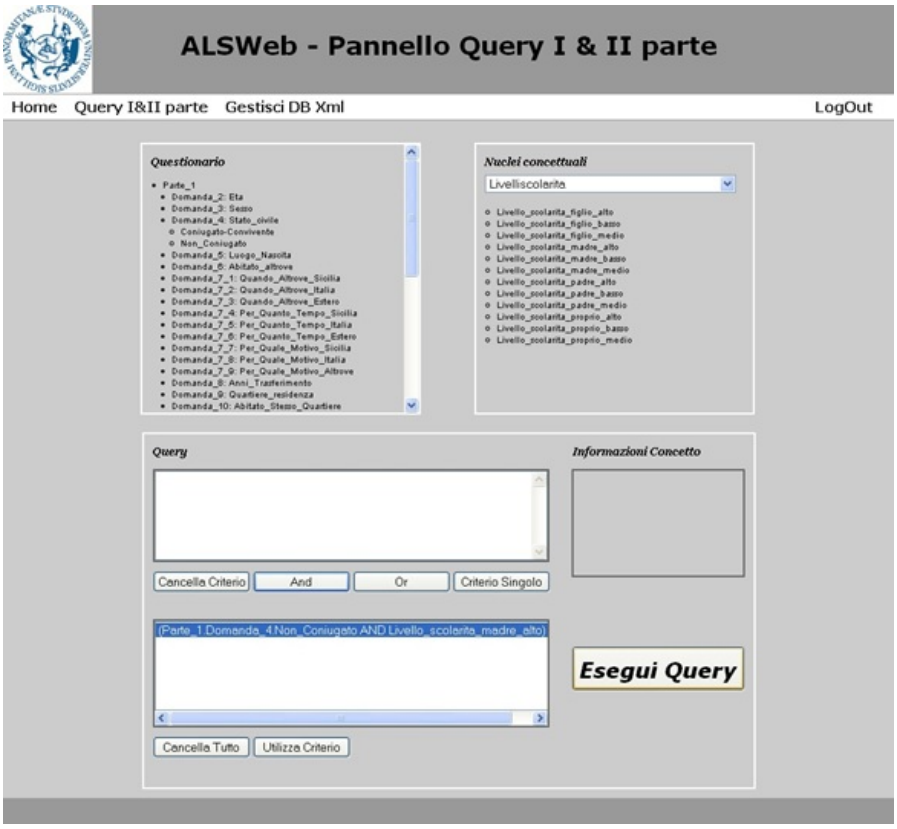

Figure 3.4: The web interface for the query composition of relational data

notation labels) about the different levels of research investigation. A tool is used to produce transcriptions directly from listening audio. Such a tool has been realized starting from the Wavesurfer software (Sjolander and Beskow (2000)), an open source tool for sound visualization and manipulation. WaveSurfer is highly customizable and adaptable to any type of analysis and audio processing the researcher wants to perform. The original version has many different configurations, each of which changes the interface by showing the most frequently used tools for any particular field of interest.

To customize WaveSurfer to ALS transcription needs, a new configuration called "ALS Transcriptio" has been created (Fig. 3.5), which consists of many panels. The first one displays the spectrogram of the audio file to be annotated. A spectrogram is a graphical representation of the sound intensity as a function of time and frequency. The second panel shows the waveform of the signal that is a graphical representation of the sound levels, and can be used as a reference for the timing labels positioning. Another panel displays a time scale that acts as a reference for all other panels, which are aligned to it. The last two panels are

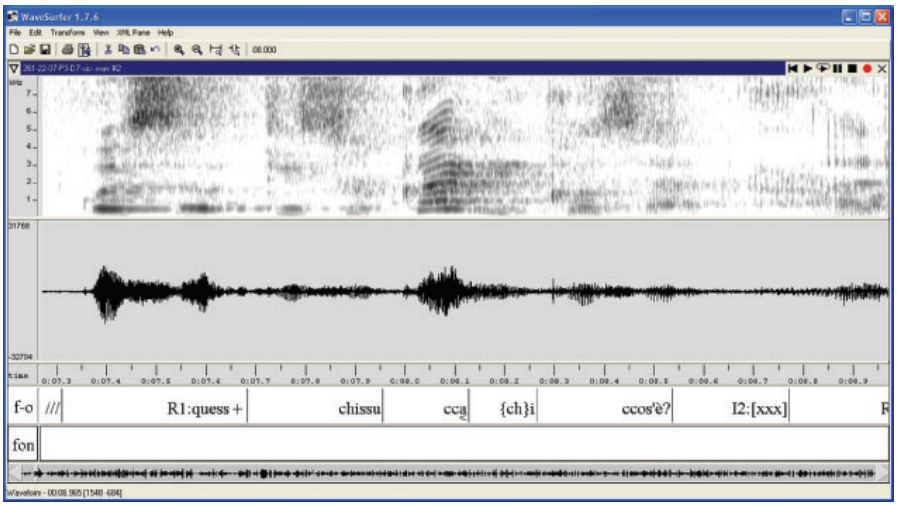

Figure 3.5: The implemented ALS transcription panels on WaveSurfer.

specifically designed for creating and managing phono-orthographic and phonetic transcriptions.

Using this customized version of WaveSurfer, the linguistic researchers can make a careful transcription of interviews. They mark a temporal interval, specifying timing labels inside one of last two panels described above, and type the related text. To save transcription, the software creates a single text file in which the start time, the final instant and the text for each label are stored.

To annotate linguistic phenomena, the domain expert works directly on the XML files that are generated from the WaveSurfer transcription, using a wrapper. This tool transforms timing labels and transcript text into valid XML files. It also add to the document some information: interviewed biographical data, reference to recorded interviewee sound file, pointer to ALS questionnaire and others. The software shows a small form window with a set of input fields. The most part of the fields are automatically filled by the wrapper, taking the information stored in both the relational database and some configuration files. This process adds structure to data that originally are full unstructured. Its output is an Xml transcription file, that isn't yet ready to be elaborated from ALSML Component.

Now the researcher can perform the annotation. He can use one of the defined annotation levels, simply associating the correct schema to XML documents produced by wrapper. So he can label the text using the tags set and related attributes, defined into the chosen XML schema. To help this operation, researchers use a XML Editor software, which makes it easier to insert annotation tags with auto-completion of code. The annotated document is then saved by the user as a file that will be uploaded successively into the ALSML Component.

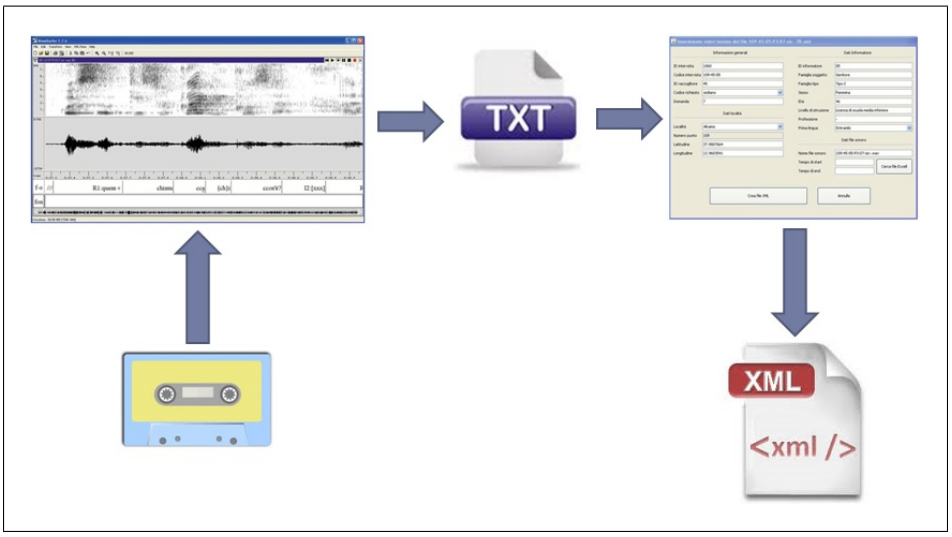

Figure 3.6: Workflow to build Xml transcription files

3.2.3 The ALSML Component

The ALSML component is deputed to manage the XML annotated documents. The system interfaces with the Oracle Berkeley DB XML (Begin et al. (2007a)) to store XML files of transcriptions. In general, a native XML database organizes files into folders that are indeed called containers. A container is the equivalent of a database for relational DBMS. To manage these containers, the XQuery (Siméon et al. (2007)) language is used. This is a programming language specified by W3C and intended to query XML documents and databases. XQuery uses the syntax of XPath expressions for selecting specific portions of XML documents, with the addition of the so-called FLWOR (For Let Where OrderBy Return) expressions to make complex queries. In particular, the FLWOR constructs played a crucial role in building the complex queries that are necessary to the investigation of linguistic phenomena.

In addition to basic functions for managing containers (i.e. add, remove, and show documents) the ALSML component provides the users with a graphical web interface to define the investigation on data. The end user can define a very complex query without writing a single line of XQuery code, but simply using the graphical interface (Fig. 3.7).

This interface is built dynamically in relation to the current annotation level that has been chosen by the linguistic researcher. All fields about the annotation tags, contained within the first two panels, are populated with the information encoded in the XML schema associated with current container. The remaining fields are populated with data contained in the headers of XML documents (see

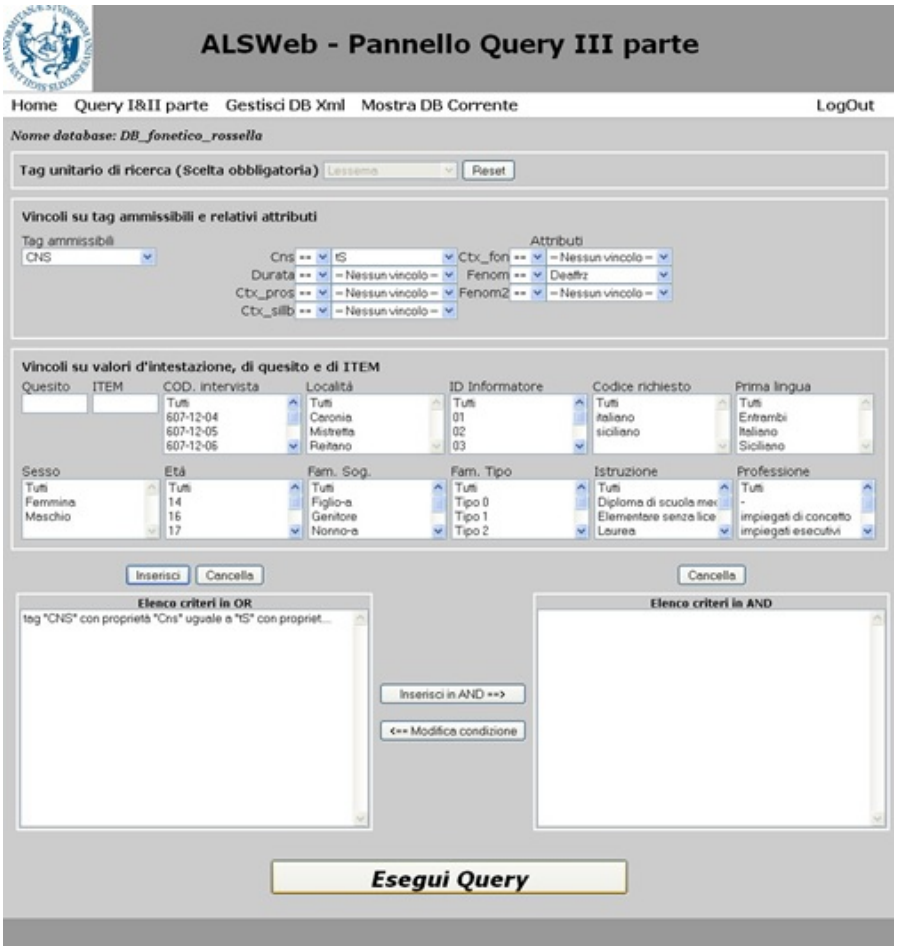

Figure 3.7: The graphical interface of the ALSML component of the framework.

paragraph 3.2.2).

By selecting the values of fields presented in the GUI, the user can define an interrogation, which will become a query to be submitted to the system. To this purpose, the first operation for the researcher is determining the so-called "unitary tag of research" among all possible linguistic tags. This tag represents the entity that the user is searching for, as the smallest piece of information to be detected. Only after this choice, the user can define constraints that characterize her research. These constraints can refer to tags (and their attributes) that have a direct lineage with the "unitary tag of research" including the unitary tag itself; otherwise the result of the query would be null. At a technical level, the choice of this tag represents a crucial phase in the composition of XQuery.

As stated previously, Xquery consists of different FLWOR nested constructs (Fig. 3.8). The degree of relationship, between the "unitary tag of research" and

another tag belongs to a constraint, and determines to which level of the FLWOR expression the same constraint must be locked.

```
for $doc in collection("file_DB/fonetica_rosa")
let $annotazione := $doc//annotazione
let $INTESTAZIONE := $annotazione/INTESTAZIONE
return
    for $frase in $annotazione//ITEM
    return
        for $tagUnitario in $frase//Lessema
        where ($tagUnitario//CNS[@Cns="tS" and @Fenom="Deaffrz"]) and ($tagUnitario[@altro_cod="sic"])
        return
        <radice xmlns:xsi='http://www.w3.org/2001/XMLSchema-instance'>
            <id nome_file="" id_intervista="{data($INTESTAZIONE/ID_intervista)}"
cod_intervista="{data($INTESTAZIONE/COD_intervista)}" localita="{data($INTESTAZIONE/Localita/Nome)}"
fam_sog="{data($INTESTAZIONE/fam_sog)}" fam_tipo="{data($INTESTAZIONE/fam_tipo)}"
sesso="{data($INTESTAZIONE/sesso)}" eta="{data($INTESTAZIONE/eta)}"
livello_istruzione="{data($INTESTAZIONE/livello_istruzione)}" professione="{data($INTESTAZIONE/professione)}"
audio_nome="{data($INTESTAZIONE/File_sonoro/nome)}" audio_tsstart="{data($INTESTAZIONE/File_sonoro/tsstart)}"
audio_tsend="{data($INTESTAZIONE/File_sonoro/tsend)}" ITEM_n="{data($frase/@n)}"
domanda="{data($INTESTAZIONE/Domanda)}" cd_richiesto="{data($INTESTAZIONE/Codice_richiesto)}"
CNS="{distinct-values($frase//CNS)[1]}" Lessema="{distinct-values($frase//Lessema)[1]}"
ITEM_tdstart="{data($tagUnitario/ancestor-or-self::*[@tdstart][last()]/@tdstart)}"
ITEM_tdend="{data($tagUnitario/ancestor-or-self::*[@tdstart][last()]/@tdend)}" />
            {$tagUnitario}
        </radice>
```

Figure 3.8: An example of an automatically built FLWOR nested construct.

After selecting this important tag, the user can compose complex search criteria using boolean logic. To this aim, the panel has two panes: the OR pane and the AND pane. Such panes are used to enter the search parameters to compose the query. The user can add multiple criteria to the OR panel. Such OR sets may be connected in a AND relationship with other ones that are in the AND pane by clicking on the central insert button. Furthermore, the user can edit a criterion that is in the AND pane moving it to OR pane through the "edit condition" button. All of these features (creation of the criteria and generation of the corresponding XQuery) are handled at client side, using JavaScript, to relieve the web server of computational burden produced by such a work. The interface detailed above gives the user great versatility in query construction; the system can generate very complex FLWOR structures. All the process is transparent to the user who does not know the complexity of XQuery generated by the system.

After defining the criteria, the system queries the Berkeley DB and displays results in a new window. As specified above, such a window (Fig. 3.9) contains

Figure 3.9: The results of an interrogation process and the option to save a new named concept

the list of tags that match the search criteria specified by user. This set of tags is divided into groups; each of them is accompanied by a small header summarizing the information about the informant which heads the tags group. Another peculiarity of this page is the presence of an audio player to listen the audio registration of an interview or a small portion of this. The result window contains also statistic information about the research, useful to linguistic researchers for their investigations.

The ALSML Component allows to increment the knowledge base of the system because it provides the possibility to package the query results into a concept that can be utilized to query the relational database. After saving the concept, it's possible to find it in the query interface of ALSDB component (Fig. 3.4) where it can be mixed with other concepts to generate a new one.

This feature allows to correlate the investigation variables belonging to first, second and third section of the ALS questionnaire. In this way, the system allows information exchanging about the saved concepts in order to build a common knowledge base regardless the creation process used by the researchers. Moreover, reasoning at different levels of abstraction is enabled. The process is intentionally managed at the user level to maintain a requisite of the project: the researchers want to have direct control and access to data, and want to be guided by their analysis and aggregation processes. Stored Concepts can be used freely if they are compliant to new findings.

3.2.4 The ALSGIS Component

The visualization tools are a relevant aspect of the project to support decisions and disseminate results. Information visualization is a very important key to achieve better research results. As stated in (McCormick B.H. (1987)) the visualization of information can be seen as a method of computing because it enables researchers to see the unseen and enrich the process of scientific discovery. Nowadays Geographic Information Systems (GISs) are increasingly being used for effective accessibility to spatial data (Jing et al. (2008)).

The ALSGIS is the part of the project related to tools and methodology able to produce results geographically referenced. One of the major factors is the geographic dependence in data evolution. To this aim a complex set of tools has been developed: some tools are stand-alone while others are collaborative and web-based. As previously stated, data exploration can be performed following several directions of investigation and different modalities. This has lead to a methodology where the interaction with users is the main driver of the process.

There are two main types of outputs for the process: maps and data sheets. The researchers have to build their own data visualization that is strictly related to the type of research to perform. A general paradigm of data visualization through visual tools defines different levels of exploration: usually users are focused on a general overview of the over-all information, then they process the data with proper filters and zooming functions and at last they ask for more details. This organization is also known as the Shneiderman's Visual Information Seeking Mantra : "overview first, zoom and filter, details-on-demand" (Shneiderman (1998)). This approach is particularly suited to obtain useful data organization and visualization.

To obtain a set of geographically referenced visualization many tools that are freely available have been personalized. A first experience (see figure 3.10) has been performed with the combination of Google Maps (gma) and Google Charts (gch).

The visualization of raw data in data sheets has to face the problem of data

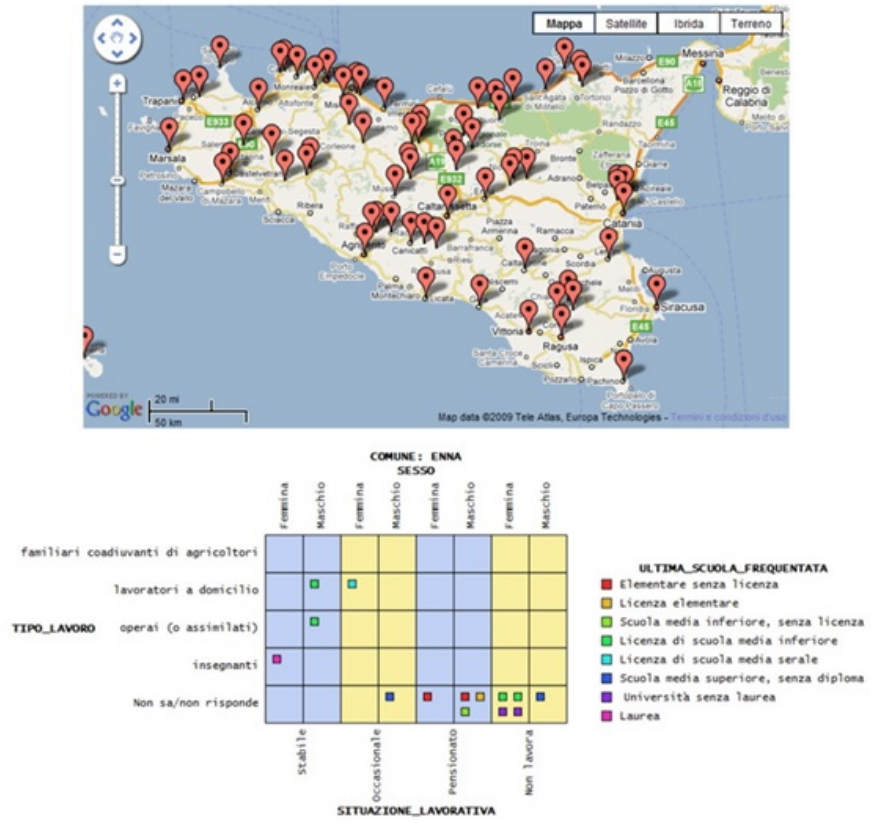

Figure 3.10: A visualization of a concept with several attributes

aggregation at run-time. The framework provides different tools to perform fast and reliable data visualization like the possibility to incorporate visualizations such as the ones defined in the Many Eyes tool. (man).

3.3 Case Study

This section reports a case study where all the features in the framework are involved. It's used to better explain how the framework allows collaborative work between users. A typical scenario involves different actors, and different instances of the same actor. The interaction between different users is managed by the framework that exposes the new concepts defined from a single researcher.

One of the main advantages of the process is that shared concepts defined by a single researcher are shared with the empirical evidences related to her investigation (that is the underlying data). In this way, the story of producing a concept is presented immediately to the other researchers.

The typical user is an expert in the sociolinguistic field or has a particular expertise in one of the annotation levels. The figure (Fig. 3.11) illustrates the main scenarios as well as the relations between different users.

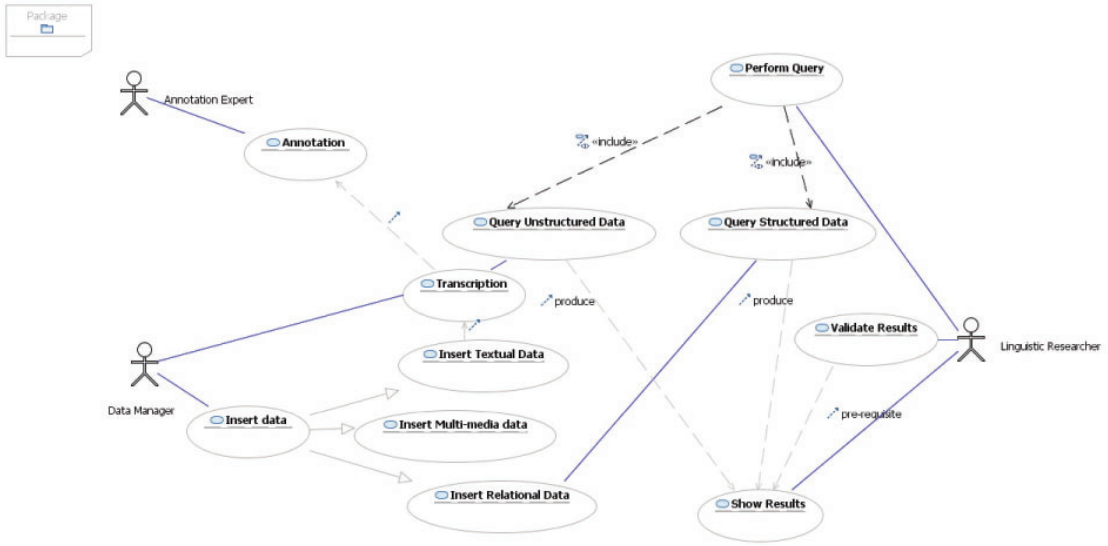

Figure 3.11: The high level scenarios definition and their correlation with different users

The researchers usually perform query operations and validate results either in a direct way by means of observation of the results or in an indirect way via the definition of new concepts in the framework. The definition of new concepts leads to new knowledge in the framework that other users have to validate in order to be employed. This validation process has been intentionally left to single users. In this way, reasoning about new concepts takes into account two main variables: the user, which creates a new concept, and the level of granularity of the concept itself. Concept definition is related to the family of saved concept via a direct descent relation. Each new concept is an offspring of another one that has been defined previously. As a consequence, it is always possible to determine how a new concept has been built.

This case study reports the creation of a concept related to the interviewee self-perception defined according to the way he felt his spoken language. Such a concept represents a particular subclass of a linguistic variable known as *diatopia*, which is the correlation between linguistic variations and the geographic area of origin. The presented question is intended to specify how the user perceives himself as different in relation to language traits in neighbor towns such as the intonation of spoken words, the phonetic aspects of pronunciation, and other relevant aspects. The process of data acquisition is performed through the ALSML component of

the infrastructure with a new level of annotation that is strictly related to the researcher, which is able to categorize the answers of users according to the main classes that have been defined.

In the presented scenario, the researcher was able to define seven different main classes with six/seven subclasses for each class. The first step of the process is the observation of the raw results like the number of occurrences of each class according to the defined classification. After the first analysis, which is useful to evaluate the classification ontology empirically, the researcher starts to search new possible correlations with other classifications defined from different users. This process is performed using the ALSDB component of the system which incorporates the quantitative results of previous observations. The researcher can use either all the classification defined from other researchers or only one aspect of the classification. This is an important side effect of the classification process, which builds close classes in the population for any given classification. As an example, in the presentation scenario the researcher can relate his classification not only to usual built-in variables like gender or level of education, but also with entirely different domains like the domains that have been defined earlier that regard the degree of formalism in the spoken language, through the definition of public and private language differences used for other researches.

This process can be iterated over different qualitative and/or quantitative variables. The result is a different classification of the same population that can be visualized in the cartography component of the framework. In this way, the process of concepts definition can be performed in a collaborative environment with an high degree of interaction with the users and the framework. The overall process has been proved to be useful by the involved research community as a non invasive way to add and formalize knowledge to allow the growth of the collective intelligence in a common environment.

3.4 Acknowledgments

This work has been partially funded under the Programma di Rilevante Interesse Nazionale (PRIN) 2009, entitled: "Atlante Linguistico della Sicilia: variazione linguistica e etnodialettologia fra nuovi modelli e nuove aree di investigazione" Authors would like to thank the ALS development team lead by prof. Giovanni Ruffino, and are especially indebted with Prof. Maria D'Agostino, Prof. Marina Castiglione, Dr. Luisa Amenta, Dr. Vito Matranga, Dr. Roberto Sottile, Dr. Giuseppe Paternostro and the many Ph.D. students at ALS for the extenuating and thoughtful discussions on the definition of the framework.

Chapter 4

Conclusions and Future Works

The work presented in this thesis is focused on the analysis of the processes responsible for the generation of Collective Intelligence (CI), i.e. the knowledge resulting from the collaboration of many individuals mediated by an intelligent system, so called Collective Knowledge Systems (CKS). The keys to getting the most from CKS, toward true collective intelligence, are tightly integrating user-contributed content and machine-gathered data, and harvesting the knowledge from this combination of unstructured and structured information.

The Emergent knowledge is the peculiar feature of collective knowledge system. To move closer to Emergent Knowledge, it is necessary to enrich user contributed with structured data. For example Xml and Semantic Web technologies can be used to add structured data to the user content. Moreover by combining structured and unstructured data it could provide a substrate for the discovery of new knowledge that is not contained in any one source.

A crucial aspect of Collective Knowledge Systems is the way people interact in the environment to communicate and to exchange information. Today human interaction is largely indirect and mediated by an increasingly wide range of technologies and devices. This new way of people interaction lead to a new research field called Human to Human Interaction (HHI). HHI is a challenging new research field where networked information systems and intelligent environments converge for the purpose of help people to cooperate for any task, any time and any where. HHI can be seen as a human to human interaction mediated by computers, in the way that machines are the more transparent as possible. Therefore the direction of the interactions is human-(to-computer)-to-human interaction. All this peculiarities allows easier knowledge exchange between people and machines, increasing the effectiveness of Collective Knowledge Systems.

The principles and best practices of Collective Intelligence to produce intelligent systems have been used in the definition of a framework for HHI in the linguistics fields, the ALS framework. The definition of an integrated methodology able to

perform a comprehensive analysis in the different fields of sociolinguistics is the ultimate goal of the Linguistic Atlas of Sicily project. The framework is intended to allow exchange of information and interoperability between users in the field of sociolinguistics. Moreover it enables the emergence of new knowledge by the synergy between the users and the system.

One of the main advantages for users is the transparency of the processes related to information retrieval and fusion starting from heterogeneous data. The framework is able to produce and process structured, semi-structured, and unstructured data, according to users' needs. An important feature is the control of data source that the users have. In this way the researchers involved in the processes can always perform check and verification of the data, underlying the concepts definition. Another important feature is the possibility to use different concepts derived from different users: this allows cross-validation and creation of new knowledge.

In the following there is a list of the main CI principles that have been applied to the ALS framework:

1. **User-generated content**. The knowledge base of the system is provided by humans participating in a social process with the purpose of sharing knowledge. The content of the ALS project are directly created by linguistic researchers that are the experts in the domain field. So the user has the complete "control of information", even some process of handling content (transcription and annotation tasks) are time consuming operations. Moreover, due to the highly interdisciplinary nature of the project, the user can generate contents at different levels: interview, transcription files, annotation files, multimedia resources and hypothesis in the form of concepts. The flow control in the user-generated content creation is one of the main CI principles adopted.

2. **Human-machine synergy.** The combination of human and machine provides a capacity to generate useful information that could not be obtained otherwise. People are the producers and customers: they are the source of knowledge, and they have real world problems and interests. Machines are the enablers: they store and remember data, search and combine data, and draw mathematical and logical inferences. In the ALS project data are organized through different modalities, from structured to unstructured, and their retrieval is performed through metadata information that are stored in the proper format. The query process is a perfect example of human-machine synergy: users don't have to possess skills in query language formulation, because the system provides an engine that allows query composition and execution only by graphical objects utilization.

3. **Increasing returns with scale.** As more people contribute, the system gets more useful. The system of rewards that attracts contributors and the computation over their contributions is stable as the volume increases. In the system the definition of a procedure able to combine results from different users allows new researcher to incorporate high value information with low effort. This feature is highly welcomed from new researchers and to pass information on the fly in order to be validated.

4. **Emergent knowledge.** The system enables computation and inference over the collected information, leading to answers, discoveries, or other results that are not found in the human contributions. The definition of emergent knowledge is the result of the search component of the system. The engine to produce knowledge is useful to evaluate hypothesis. Only the user knows if the obtained result is compliant with the hypothesis, so the researchers decides if the concept has to be added to the knowledge base. The emergent knowledge produced is directly bounded to data through the query composition process.

5. **Need for structured data** To move closer to Emergent Knowledge, it is necessary to enrich user-contributed data with structured data. The framework is able to incorporate high level concepts into semi-structured data through proper correspondences. The level of granularity of the correspondence reflects the grain of the concepts. The more the concept is specific the less the data related are in number. This mechanism is completely transparent to user that can interface with the system without knowledge of the data aggregation engine. Moreover some tools for data preparation (transcription and annotation) are precisely designed to add structure to data using xml-like techniques. It means that this information can be automatically used by the system to generate new knowledge.

Future works are in the direction of a more explicit definition of the process of knowledge sharing with the introduction of a search engine service. This service will be used from researchers to address particular findings that are relevant to them or to other users. The preliminary effort is the definition of a fair trade-off between the increasing of computational load for the users who has to annotate somehow the created concepts and the effective benefits of new and automatic concepts highlighted by the search engine.

We must recognize that intelligence is distributed wherever there is humanity, and that this intelligence, distributed everywhere, can be exploited to the full by the new technologies. Today, if two people know two things away complementary, through internet and Web 2.0, they can really get into communication with each

other, share their knowledge, to cooperate. Resolution of problems through collaboration and sharing is the basic principle of collective intelligence. On one side is accentuated individuality, each user in fact tends to assert itself and to want to be different from the crowd at all costs, on the other hand it is inevitable to become part of a group. Social networks and social media in general are strongly influenced by interactions between people.

Dialogue, discussion, complaints or simply highlighting certain content: everything is expanded through technology, characterized mainly by the rate of diffusion and the multiple modes of communication. Internet's users often become generators of content. Blogs are an expression pattern of their skills, their ideas, or their interests. YouTube is a prime example of how the users of the internet they are also producer. But, individuality, can also give rise to collaboration. It is possible to say that with social networks has opened the era of sharing knowledge and information.

Appendix A

A platforms for HHI in Large Social Events

In this appendix an information system built to provide people with rich user experiences, when attending museums or exhibits, is analyzed according to the human-to-human interaction (HHI) research domain. The HHI (see section 2.3) is a challenging new domain, an exciting field of design originating from the convergence of a few well-established research areas, such as traditional graphical user interfaces (GUIs), tangible user interfaces (TUIs), touchless gesture user interfaces (TGUIs), voice user interfaces (VUIs), and brain computer interfaces (BCIs) (Gentile et al. (2011)).

Well-studied technologies as ubiquitous and pervasive computing currently implement and employ networked information systems on many important applications. In addition, the ever-growing communication bandwidth offers new opportunities for bringing richer multimedia contents readily available in mobile devices and public interactive system such as large-scale displays. While large-scale displays keep popping up in public places, but very few of them at present show significant interaction capabilities. In this panorama, cloud-based information systems and intelligent environments surrounding people do converge for the purpose of better satisfaction of users' requirements and anticipation of their needs.

In the recent years, many HCI and Interaction Designer researchers carried out a significant number of studies on interaction techniques in large spaces. In (Ruan et al. (2010)) Ruan et al. present TouchInteract: a novel interaction technique with large displays using touchscreen phones, allowing users to remotely control the large displays using their touchscreen on mobile phone, allowing and enhancing multiuser (humanhuman) interaction in large open space. In (Mladenov et al. (2012)) authors propose a system which aims at making brain-computer interfaces popular with consumer products, providing a more natural human computer interaction (HCI). They aspire to make the BCI systems more attractive to users

and to develop applications that go beyond the medical care services. In (Vlaming et al. (2008)) authors point out how multi-touch interaction with small and large displays has received much attention in recent years. They set up a system that tracks the thumb and index fingers per hand for both motion and grabbing gestures detection. This technique allows users to indirectly interact with objects on large presentation screens from a distance, possibly with the help of a separate small display, enhancing human-human interaction in large spaces.

In this chapter is presented QRouteMe, that could be considered as a virtual place built upon a real trade show implementing social networks capabilities, in which visitors and exhibitors meet together to communicate, share and discuss about products, experiences, and possible trade exchanges. To realize these goals it is essential to provide users with attractive interfaces to satisfy their needs through an enjoyable user experience. The social software definition encompasses a set of software components that allow users to interact and share data. This system has been built and deployed to address some specific issues in the field of Human to Human Interaction. Here the design and the evolution of the QRouteMe platform from an architectural point of view is discussed, showing several case studies.

A.1 Related work

The definition and implementation of complex systems able to support users in indoor environments like an exhibit or a museum is an active and multidisciplinary research field. This issue has been researched from many perspectives starting from the process of contents definition and organization to personalization from different users or localization for indoor environments.

An example of complex system is the Cyberguide (Genco et al. (2006)) work, which defines a set of different prototypes both for indoor and outdoor guides. The system was designed as a combination of four main components: a cartographer component including the map (or maps) of the physical environments that the tourist is visiting, a librarian component which is the information repository containing all the information to be presented, a navigator component used to keep track of the users' positions in the environment, and a messenger component used to record message exchanges to/from users to system.

The Hippie/HIPS project (Oppermann and Specht (2000)) is another relevant system that is focused on development of an exhibition guide, providing guidance and information services. Starting from the observations about the visitor's movements through the exhibition the system create a user profile and suggests other interesting exhibits or paths inside the current exhibits.

In the PEACH project (Stock et al. (2007)) the generation of some position related contents and post-visit reports are automatically performed. The CHIP

project (Aroyo et al. (2007)) tries to combine Semantic Web techniques to provide personalized access to digital museum collections both online and in the physical museum. The above-mentioned projects and other related work try to exploit the possibility of automatically define related information for a guide. Most of the works are focused on an explicit definition of a knowledge base while some works tries to implicitly define a user model. The user model definition is based on statistical models rather than recommendation techniques (Albrecht and Zukerman (2007)).

Another point of view to build a museum guide is to target not just a single user but also a group visiting a museum. The Sottovoce (Aoki et al. (2002)) system is designed specifically for this goal, providing for a communication mechanism to support interaction. The system provides user with tools for human-human interaction during the visit.

The user position is another important asset for these systems that is addressed distinguishing between indoor positioning and outdoor positioning. Both technologies require an electronic infrastructure to register measurements. In Mulloni et al. (2009) the Signpost system is presented that is used as a location-based conference guide. The system works only with Windows Mobile phones but is able to be used in largescale events. Another way to achieve the same functionality is through the detection of the position by comparing a set of floorplans against an image taken from the cell-phone camera (Hile and Borriello (2008)). This method has a major disadvantage because it requires that all the floorplans for a particular building be available and processed ahead of use.

A.2 QRouteMe

QRouteMe is an information system for exhibition spaces implementable in fairs, exhibits, and large events. The system is made for large display areas or places that receive large crowds, needing interactive information. The system tasks range from the positioning of users within the site to the trade promotion.

QRouteMe is a system that integrates totem/kiosks and mobile devices for the information fruition in display areas. Through these tools, users (tourists, visitors, business men) can access the information (text, images, videos, views) provided by system, identify their own position inside the exhibition area and receive related information, locate points of interest and directions to reach them. Furthermore it is possible to visualize specific augmented reality views of objects, artifacts, works of art, products, or, more generally, any item of interest.

On the other side, QRouteMe administrators can perform a number of tasks, such as monitoring the distribution of users, accessing to back-end information in real time, updating information about the POIs to reflect last-minute changes,

displaying usage statistics in real time, and providing for event information to the public with ad-hoc viewgraphs projected on suitable spots (an example is given in Figure A.1).

QRouteMe establishes a permanent link with the users of the event that remains active even after the end of the event thanks to the mobile application and the mobile information site, both important tools to maintain contact with the users all year around.

A.2.1 System Architecture

QRouteMe is a client-server solution composed by tree subsystems: QRouteMe Platform, QRouteMe On Site and QRouteMe Front End (see figure A.1).

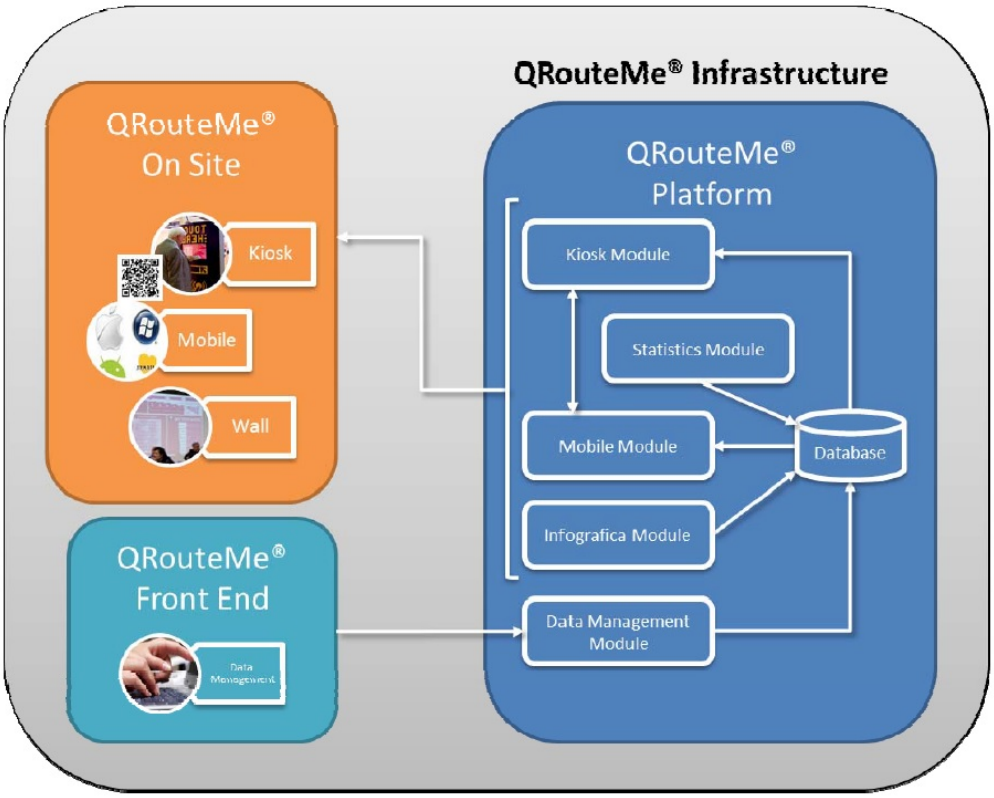

Figure A.1: QRouteMe infrastructure

The first of these subsystems implements the majority of the QRouteMe functions and acts as the server side part of the system. It includes different modules and the system database, which stores the information domain. QRouteMe Platform contains the Mobile and Kiosk modules that manage the information to be

sent to different devices, that interact with the system and, if necessary, synchronize the information on the various devices. These two modules also update into the database all user access information that is managed by Statistics and "Infographics" modules.

The Statistics module collects data on system access and usage, and organizes it based on organizers specific requests (number of accesses, access to certain information, website real time reports, event notifications and any useful information).

The "infographics" module present the gatherd data to the public, according to project requirements. QRouteMe On Site is deputed to manage the user interface of the system, designed to ensure a good user interaction. It implements all the modules that manage the touch screen totem graphical user interface, the infographics screen projection into the large screens on the wall and the mobile site of QRouteMe interface that provides a dedicated web site for all type of smartphones and a mobile app specifically designed for Apple devices.

QRouteMe Front End incorporates also a subsystem allowing the customer to create, read, modify and delete the data that the system shows to users.

A.2.2 Scenarios

To better understand the features of QRouteMe is useful to give some use cases of the system. A possible scenario is the description of the experience of a user that, inside an exhibit, arrives in front of a interactive touch screen totem and, through its interface, navigates the content provided by the system, until he identifies an exhibitor, whose stand he wants to visit. Once he has obtained the page, the system displays the map of the pavilion and the path between the selected exhibitor's stand and the totem utilized by the user. In addition the system offers the same path through a QR (Quick Response) code displayed on the screen; the visitor equipped with a camera smartphone, can point his mobile device on the QR code, which provides immediate access to the mobile version of the map with the same route displayed on the totem screen, so that he can take with him during the visit inside the stand.

Another possible scenario involves a mobile user who downloads the app on his smartphone to browse the information provided by QRouteMe. He looks for the exhibitor whose stand he wants to visit and so he select the features "found path". At this point, the application asks to the visitor to scan the QR code closest to its current location. Once decoded the QR code, the system shows on smartphone screen a map of the pavilion in which it highlighted the exhibitor stand and the path between this and the point where the visitor is.

A.3 Implementations

A.3.1 Vinitaly 2011 in Verona, Italy

The QRouteMe system has been implemented in several exhibitions. Its debut was coming to Verona in April 2011 on the occasion of the Vinitaly 2011. This implementation, called "Sicilia@Vinitaly2011", was performed within the pavilion used by wine producers that join in the "Istituto Regionale della Vite e del Vino della Regione Sicilia", the Sicily's Institute for the promotion of regional wines and producers. In this pavilion of 8,000 square meters, 14 fixed information points were installed with an ad hoc interface for the fruition of information about the Sicilian wineries and their wines.

The system contained information on 234 exhibitors for each of which has created a QR to direct access to their information sheet. The system provided free Wifi to its users and also included a mobile website for smartphone and a devices application ("app") for mobile devices with the operating system iOS (iPhone, iPod, iPad). Sicilia@Vinitaly2011 recorded, during the five-day of fair, around 40,000 page view, 5,000 Wifi access, 500 iPhone App distributed. This app is still available on the Apple Store and with more than 2,000 downloads to date.

A.3.2 London International Wine Fair 2011 in London, UK

For the second time QRouteMe was implemented in London International Wine Fair in May 2011, a big international wine fair. An area of 400 sq. mt. in which 25 Sicilian wineries exposed their product, in which we deployed 3 touch screen totems. This was a small implementation of Sicilia@Vinitaly2011 that represent a simplification of the QRouteMe system. A peculiarity of this implementation was the building of a section to navigate the exhaustive information about the Sicilian wine that the visitors of fair could taste at the tasting desk.

A.3.3 Fruit Logistica 2012 in Berlin, Germany

Another application of QRouteMe was implemented in Berlin in February 2012 in occasion of Fruit Logistica 2012. This is the world's leading trade fair for the fresh fruit and vegetable business to which 20 Sicilian agricultural producers exposed own product. In this occasion, inside an area of 400 sq. mt., the users had access to information, by means of both 2 kiosks (interactive totems), 10 iPad (tablet PCs), suitably placed close to the exhibitors booths, and through their own smartphones, if equipped with the needed hardware/software features. Even in this case we deployed a website formatted for presentation of contents on mobile

devices and an application ("app") for iOS mobile devices, made available for free at the Apple Store.

The self-location task on site has been accomplished by kiosks (totem) showing on the screen the highlighted kiosk position on the stand plan.

The exhibitors information data have been made available directly, using QR codes (quick response code) assigned to each exhibitor, and integrated with the printed graphics of each booth. In addition, a video trailer (duration 1 minute) was created and installed on the 2 totems to indicate its interactivity.

In this implementation the use of system was more focused on the iPad App than on the totems, because the producers used the tablets like a tool for their product promotion to the fair's visitors. This app is still available on the Apple Store and it has been downloaded about 2,000 times from all around the world so far.

A.3.4 ProWein 2012 in Dusseldorf, Germany

QRouteMe was implemented on March 2012 in Dusseldorf in occasion of Prowein 2012, that is considered the world leading trade fair for the wine and spirits industry. Similarly to the London International Wine Fair, this was a small implementation of Sicilia@Vinitaly2011, with three fixed information kiosks within an area of about 400 sq. mt. and 30 Sicilian exhibitors participating to the event.

A.3.5 Vinitaly 2012 in Verona, Italy

The last QRouteMe implementation was putted inside a bigger project called SiciliaWineCloud, that was built in occasion of Vinitaly 2012, the bigger Italian exhibit of wine that is held every year in Verona.

WineCloud is an integrated project of communications and promotions dedicated to the Wines of Sicily. WineCloud involves the building of an application for mobile devices and another one for kiosks (upgradeable of all the other events organized by the customer). That system allows the connection of the main actors in the promotion like the IRVOS (Istituto Regionale della Vite e del Vino - Regione Sicilia), wineries and wine world enthusiasts, in order to achieve a virtuous cycle of involvement with the following communicational objectives: to put Sicilian wine in the center of attention of promotional word and to allow involvement of communication actors for 365 days a year, considering the wine exhibit like a great moments of focus and launch for the system.

From the other side, the benefits to the customer are manifold and include the following availability:

- A virtual fair service, where the actors are present round-the-year;

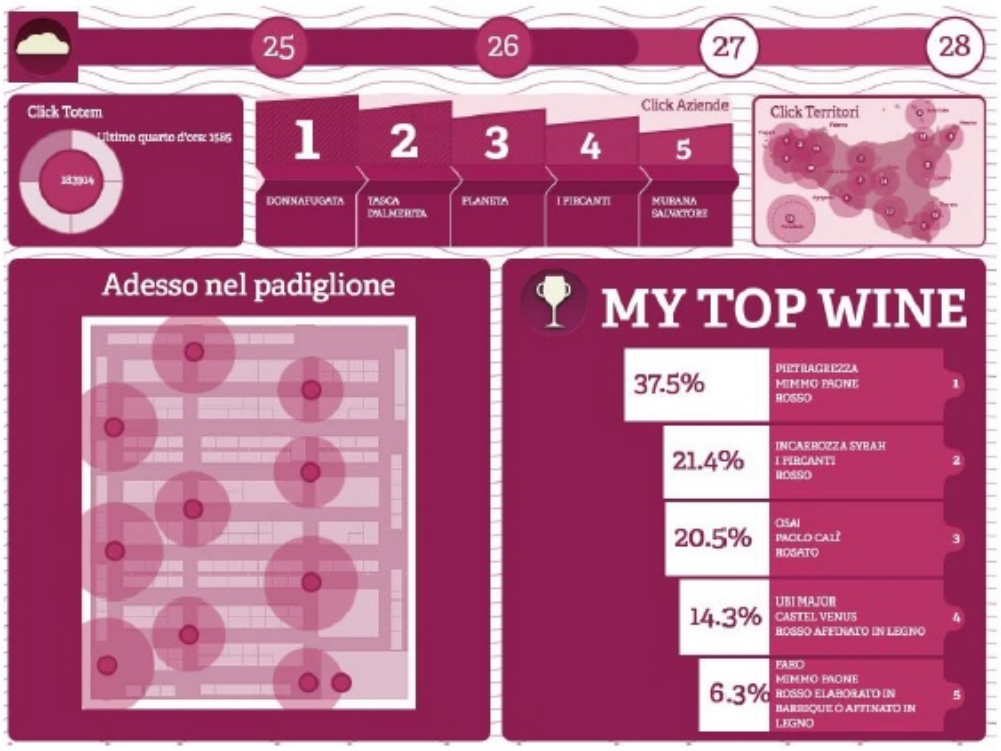

Figure A.2: Real time infographic

- An high value added services to be provided to wine producers (QR certified promocommunicational additional space, access by businesses to the user base acquired);
- A system of promotional communication for institutional purpose covering all major Internet channels (web and social networks).

Similarly to Sicilia@Vinitaly2011, the SiciliaWineCloud application for Vinitaly 2012 consisted of 14 information points (touch screen totem), fixed inside a pavilion of 8,000 sm., with ad-hoc software for the information fruition, 187 exhibitors for which he was created a QR for direct access to information in context, free WiFi, a website formatted for presentation of content on mobile devices, an application ("app") for mobile devices with iOs operating system (iPhone, iPod, iPad) and the construction of a back-end and front-end platform to manage both the information about the wineries and the promotional advises that was displayed inside screen.

Compared to other QRouteMe applications, SiciliaWineCloud has seen the introduction of some innovations. The first was the creation of an interactive social game, called "My Top Wine", that provided the possibility to rate the wines

tasted by mobile users. During the four days of exhibit, the system rewards the best wine and one of users who voted for him.

The second innovation consisted of a real-time "infographic" dashboard projected in the Business Area, that showed the real time totem usage and wines votes (Figure A.2). These two innovations have made the system more attractive to the user, who has been involved in the first person to express their preferences about wines. The social aspect of SiciliaWineCloud has increased the usage statistics of the system compared to those recorded the previous year to the same event.

Bibliography

Mit center for collective intelligence. URL http://cci.mit.edu/.

Google Chart API. URL https://developers.google.com/chart/.

Google Map API. URL https://developers.google.com/maps/?hl=it.

Many eyes. data visualisation tools from ibm. URL http://www-958.ibm.com/software/data/cognos/manyeyes/.

The sax project home page. URL http://www.saxproject.org/.

Eugene Agichtein and Luis Gravano. Snowball: Extracting relations from large plain-text collections. In *Proceedings of the Fifth ACM Conference on Digital Libraries*, DL '00, pages 85–94, New York, NY, USA, 2000. ACM. ISBN 1-58113-231-X. doi: 10.1145/336597.336644. URL http://doi.acm.org/10.1145/336597.336644.

Luis von Ahn. Games with a purpose. *Computer*, 39(6):92–94, June 2006. ISSN 0018-9162. doi: 10.1109/MC.2006.196. URL http://dx.doi.org/10.1109/MC.2006.196.

David W. Albrecht and Ingrid Zukerman. Introduction to the special issue on statistical and probabilistic methods for user modeling. *User Model. User-Adapt. Interact.*, 17(1-2):1–4, 2007. URL http://dblp.uni-trier.de/db/journals/umuai/umuai17.html#AlbrechtZ07.

Paul M. Aoki, Rebecca E. Grinter, Amy Hurst, Margaret H. Szymanski, James D. Thornton, and Allison Woodruff. Sotto voce: exploring the interplay of conversation and mobile audio spaces. In Dennis R. Wixon, editor, *CHI*, pages 431–438. ACM, 2002. ISBN 1-58113-453-3. URL http://dblp.uni-trier.de/db/conf/chi/chi2002.html#AokiGHSTW02.

C. Apte and Se June Hong. Predicting equity returns from securities data, 1995.

Lora Aroyo, Natalia Stash, Yiwen Wang, Peter Gorgels, and Lloyd Rutledge. Chip demonstrator: Semantics-driven recommendations and museum tour generation. In Karl Aberer, Key-Sun Choi, Natasha Fridman Noy, Dean Allemang, Kyung-Il Lee, Lyndon J. B. Nixon, Jennifer Golbeck, Peter Mika, Diana Maynard, Riichiro Mizoguchi, Guus Schreiber, and Philippe CudrÃ©-Mauroux, editors, *ISWC/ASWC*, volume 4825 of *Lecture Notes in Computer Science*, pages 879–886. Springer, 2007. ISBN 978-3-540-76297-3. URL `http://dblp.uni-trier.de/db/conf/semweb/iswc2007.html#AroyoSWGR07`.

X. Artola, A. D. de Ilarraza, A. Soroa, and A. Sologaistoa. Dealing with complex linguistic annotations within a language processing framework. *Audio, Speech, and Language Processing, IEEE Transactions on*, 17(5):904 –915, jul. 2009. ISSN 1558-7916. doi: 10.1109/TASL.2009.2018565.

Soren Auer and Jens Lehmann. What have innsbruck and leipzig in common? extracting semantics from wiki content. In *In ESWC*, pages 503–517. Springer, 2007.

MichÃ¨le Basseville and Igor V. Nikiforov. Detection of abrupt changes: Theory and application, 1993.

Clinton Begin, Brandon Goodin, and Larry Meadors. *Ibatis in Action*. Manning Publications Co., Greenwich, CT, USA, 2007a. ISBN 1932394826.

Clinton Begin, Brandon Goodin, and Larry Meadors. *Ibatis in Action*. Manning Publications Co., Greenwich, CT, USA, 2007b. ISBN 1932394826.

Donald J. Berndt and James Clifford. Finding patterns in time series: A dynamic programming approach. In *Advances in Knowledge Discovery and Data Mining*, pages 229–248. 1996. URL `http://dblp.uni-trier.de/db/books/collections/fayyad96.html#BerndtC96`.

Steven Bird, David Day, John Garofolo, John Henderson, Christophe Laprun, and Mark Liberman. Atlas: A flexible and extensible architecture for linguistic annotation, 2000.

Christian Bizer, Richard Cyganiak, Tobias Gaus, and Freie UniversitÃ¤t Berlin. The rdf book mashup: From web apis to a web of data, 3rd workshop on scripting for the semantic web. In *6, 2007. Available online as CEUR Workshop Proceedings, ISSN 1613-0073, online CEUR-WS.org/Vol248/paper4.pdf*, page 2007, 2007.

Jon Bosak, Tim Bray, Dan Connolly, Eve Maler, Gavin Nicol, C. Michael Sperberg-McQueen, Lauren Wood, and James Clark. Dtd (document type definition), 2000. URL `http://www.w3.org/XML/1998/06/xmlspec-report.htm`.

T. Bray, J. Paoli, C.M. Sperberg-McQueen, and E. Maler. Extensible markup language (xml), 1.0 second edition. *W3C Recommendation*, 2006. URL `http://www.w3.org/TR/2006/REC-xml-20060816/`.

D. Brickley and L. Miller. FOAF vocabulary specification. http://xmlns.com/foaf/spec/, November 2007. URL `http://xmlns.com/foaf/spec/`.

Michael J. Cafarella, Doug Downey, Stephen Soderland, and Oren Etzioni. Knowitnow: Fast, scalable information extraction from the web. In *HLT/EMNLP 2005*, 2005.

Jean Carletta, David McKelvie, and Amy Isard. Supporting linguistic annotation using xml and stylesheets. In *READINGS IN CORPUS LINGUISTIC*, 2002.

Hamish Cunningham, Yorick Wilks, and Robert J. Gaizauskas. Gate - a general architecture for text engineering. In *Proc. 16th Conf. Comput. Linguist. Assoc. Computat. Linguist.*, 1996.

Steven DeRose and James Clark. XML path language (XPath) version 1.0. November 1999. http://www.w3.org/TR/1999/REC-xpath-19991116.

P. Di Maio. Digital ecosystems, collective intelligence, ontology and the 2nd law of thermodynamics. In *Digital Ecosystems and Technologies, 2008. DEST 2008. 2nd IEEE International Conference on*, pages 144–147, 2008. doi: 10.1109/DEST.2008.4635217.

Darcy DiNucci. Fragmented future. *Print*, 53(4):32, 1999.

Denise Draper, Alon Y. HaLevy, and Daniel S. Weld. The nimble xml data integration system. In *In ICDE*, pages 155–160, 2001.

Douglas C. Engelbart. *A conceptual framework for the augmentation of man's intellect*, volume 1, pages 1–29. Spartan Books, Washington, 1963.

David C. Fallside and Priscilla Walmsley. XML schema part 0: Primer second edition. October 2004. http://www.w3.org/TR/2004/REC-xmlschema-0-20041028/.

Usama Fayyad, Usama Fayyad, Gregory Piatetsky-shapiro, Gregory Piatetsky-shapiro, Padhraic Smyth, and Padhraic Smyth. Knowledge Discovery and Data Mining: Towards a Unifying Framework. In *Proc. 2nd Int. Conf. on Knowledge Discovery and Data Mining*, pages 82–88, 1996.

Tian Feng, Han Xiao-bing, and Wu Feng-bo. The heterogeneous data integration based on xml in coal enterprise. volume 1, pages 438 –441, dec. 2008. doi: 10.1109/ISCSCT.2008.47.

Alessandro Genco, Salvatore Sorce, Giuseppe Reina, and Giuseppe Santoro. An agent-based service network for personal mobile devices. *IEEE Pervasive Computing*, 5(2):54–61, 2006. ISSN 1536-1268. doi: http://doi.ieeecomputersociety.org/10.1109/MPRV.2006.22.

Antonio Gentile, Antonella Santangelo, Salvatore Sorce, and Salvatore Vitabile. Human-to-human interfaces: emerging trends and challenges. *IJSSC*, 1(1):3–17, 2011. URL `http://dblp.uni-trier.de/db/journals/ijssc/ijssc1.html#GentileSSV11`.

C.N. Glymour. *Discovering causal structure: artificial intelligence, philosophy of science, and statistical modeling*. CMU-LCL-report. Academic Press, 1987. ISBN 9780122869617. URL `http://books.google.it/books?id=zKnuAAAAMAAJ`.

Tom Gruber. Collective knowledge systems: Where the social web meets the semantic web. In *Web Semantics: Science, Services and Agents on the World Wide Web*, 2008.

I. Guyon, N. Matic, and V. Vapnik. Discovering informative patterns and data cleaning, 1996.

David Heckerman. Bayesian networks for data mining. *Data Min. Knowl. Discov.*, 1(1):79–119, January 1997. ISSN 1384-5810. doi: 10.1023/A:1009730122752. URL `http://dx.doi.org/10.1023/A:1009730122752`.

Harlan Hile and Gaetano Borriello. Positioning and orientation in indoor environments using camera phones. *IEEE Computer Graphics and Applications*, 28 (4):32–39, 2008. URL `http://dblp.uni-trier.de/db/journals/cga/cga28.html#HileB08`.

Jeff Howe. The rise of crowdsourcing. *Wired Magazine*, 14(6), 06 2006. URL `http://www.wired.com/wired/archive/14.06/crowds.html`.

Tan Jing, Xu Juan, and Wan Li. Open source software approach for internet gis and its application. *Intelligent Information Technology Applications, 2007 Workshop on*, 3:264–268, 2008. doi: http://doi.ieeecomputersociety.org/10.1109/IITA.2008.501.

Michael Kay. Xsl transformations (xslt) version 2.0, 2007. URL `http://www.w3.org/TR/2007/REC-xslt20-20070123/`.

Hsiang Hui Lek, D.C.C. Poo, and N.K. Agarwal. Knowledge community (k-comm): Towards a digital ecosystem with collective intelligence. In *Digital Ecosystems and Technologies, 2009. DEST '09. 3rd IEEE International Conference on*, pages 211–216, 2009. doi: 10.1109/DEST.2009.5276690.

V. Lertnattee, S. Chomya, T. Theeramunkong, and V. Sornlertlamvanich. Applying collective intelligence for search improvement on thai herbal information. In *Computer and Information Technology, 2009. CIT '09. Ninth IEEE International Conference on*, volume 2, pages 178–183, 2009. doi: 10.1109/CIT.2009.52.

Dekang Lin and Patrick Pantel. Concept discovery from text. In *In Proceedings of Conference on Computational Linguistics*, pages 577–583, 2002.

Lucian Vlad Lita and Jaime G. Carbonell. Instance-based question answering: A data-driven approach. In *EMNLP*, pages 396–403. ACL, 2004. URL `http://dblp.uni-trier.de/db/conf/emnlp/emnlp2004.html#LitaC04`.

P.E. Maher and J.L. Kourik. A knowledge management system for disseminating semi-structured information in a worldwide university. In *Management of Engineering Technology, 2008. PICMET 2008. Portland International Conference on*, pages 1936–1942, 2008. doi: 10.1109/PICMET.2008.4599814.

Christopher J. Matheus, Gregory Piatetsky-shapiro, and Dwight Mcneill. 20 selecting and reporting what is interesting: The kefir application to healthcare data.

Brown M.D. McCormick B.H., DeFanti T.A. Visualization in scientific computing and computer graphics. In *ACM Special Interest Group on GRAPHics and Interactive Techniques*, 1987.

H. Mizuyama and Y. Maeda. A prediction market system using sips and generalized lmsr for collective-knowledge-based demand forecasting. In *Computers and Industrial Engineering (CIE), 2010 40th International Conference on*, pages 1–6, 2010. doi: 10.1109/ICCIE.2010.5668201.

T. Mladenov, K. Kim, and S. Nooshabadi. Accurate motor imagery based dry electrode brain-computer interface system for consumer applications. In *Consumer Electronics (ISCE), 2012 IEEE 16th International Symposium on*, pages 1–4, 2012. doi: 10.1109/ISCE.2012.6241718.

Raymond J. Mooney and Razvan Bunescu. Mining knowledge from text using information extraction. *SIGKDD Explor. Newsl.*, 7(1):3–10, 2005. ISSN 1931-0145. doi: http://doi.acm.org/10.1145/1089815.1089817. URL `http://portal.acm.org/citation.cfm?id=1089817`.

Alessandro Mulloni, Daniel Wagner, Istvan Barakonyi, and Dieter Schmalstieg. Indoor positioning and navigation with camera phones. *IEEE Pervasive Computing*, 8(2):22–31, April 2009. ISSN 1536-1268. doi: 10.1109/MPRV.2009.30. URL `http://dx.doi.org/10.1109/MPRV.2009.30`.

San Murugesan. Understanding web 2.0. *IT Professional*, 9(4):34–41, July 2007. ISSN 1520-9202. doi: 10.1109/MITP.2007.78. URL `http://dx.doi.org/10.1109/MITP.2007.78`.

Reinhard Oppermann and Marcus Specht. A context-sensitive nomadic exhibition guide. In Peter J. Thomas and Hans-Werner Gellersen, editors, *HUC*, volume 1927 of *Lecture Notes in Computer Science*, pages 127–142. Springer, 2000. ISBN 3-540-41093-7. URL `http://dblp.uni-trier.de/db/conf/huc/huc2000.html#OppermannS00`.

Tim O'Reilly. O'reilly network: What is web 2.0, September 2005. URL `http://www.oreillynet.com/lpt/a/6228`.

Patrick Pantel. Espresso: Leveraging generic patterns for automatically harvesting semantic relations, 2006.

Patrick Pantel and Deepak Ravichandran. Automatically labeling semantic classes. In *HLT-NAACL*, pages 321–328, 2004. URL `http://dblp.uni-trier.de/db/conf/naacl/naacl2004.html#PantelR04`.

Gregory Piatetsky-Shapiro. Knowledge discovery in real databases: A report on the ijcai-89 workshop. *AI Mag.*, 11(5):68–70, January 1991. ISSN 0738-4602. URL `http://dl.acm.org/citation.cfm?id=124898.124915`.

Howard Rheingold. *Smart Mobs: The Next Social Revolution*. Basic Books, October 2003. ISBN 0738208612. URL `http://www.amazon.ca/exec/obidos/redirect?tag=citeulike04-20{&}path=ASIN/0738208612`.

Hanqing Ruan, Yi Qian, Yong Zhang, and Min Zhou. Touchinteract: An interaction technique with large displays using touchscreen-phone. *Ubiquitous, Autonomic and Trusted Computing, Symposia and Workshops on*, 0:262–265, 2010. doi: http://doi.ieeecomputersociety.org/10.1109/UIC-ATC.2010.36.

Giovanni Ruffino and Mari D'Agostino. *I Rilevamenti Sociovariazionali. Linee Progettuali.* 2005. ISBN 8890214813.

Giuseppe Russo, Antonio Gentile, Roberto Pirrone, and Vincenzo Cannella. Xml-based knowledge discovery for the linguistic atlas of sicily (als) project. In Leonard Barolli, Fatos Xhafa, and Hui-Huang Hsu, editors, *CISIS*, pages 98–104. IEEE Computer Society, 2009. ISBN 978-0-7695-3575-3. URL `http://dblp.uni-trier.de/db/conf/cisis/cisis2009.html#RussoGPC09`.

Warren Sack. Discourse diagrams: Interface design for very large-scale conversations. In *HICSS*, 2000. URL `http://dblp.uni-trier.de/db/conf/hicss/hicss2000-3.html#Sack00`.

Ben Shneiderman. *Designing the User Interface - Strategies for Effective Human-Computer Interaction.* Addison-Wesley Longman, Reading, MA, 3rd edition, 1998.

B. W. Silverman. *Density Estimation for Statistics and Data Analysis.* Chapman & Hall, London, 1986.

Jérôme Siméon, Don Chamberlin, Daniela Florescu, Scott Boag, Mary F. Fernández, and Jonathan Robie. XQuery 1.0: An XML query language. January 2007. http://www.w3.org/TR/2007/REC-xquery-20070123/.

Kare Sjolander and Jonas Beskow. Wavesurfer - an open source speech tool. In *INTERSPEECH*, pages 464–467. ISCA, 2000. URL `http://dblp.uni-trier.de/db/conf/interspeech/interspeech2000.html#SjolanderB00`.

Oliviero Stock, Massimo Zancanaro, Paolo Busetta, Charles Callaway, Antonio Krüger, Michael Kruppa, Tsvi Kuflik, Elena Not, and Cesare Rocchi. Adaptive, intelligent presentation of information for the museum visitor in peach. *User Modeling and User-Adapted Interaction*, 17(3):257–304, July 2007. ISSN 0924-1868. doi: 10.1007/s11257-007-9029-6. URL `http://dx.doi.org/10.1007/s11257-007-9029-6`.

Fabian M. Suchanek, Georgiana Ifrim, and Gerhard Weikum. Combining linguistic and statistical analysis to extract relations from web documents. In *12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*

(KDD 2006), pages 712–717, New York, NY, USA, 2006. ACM. URL `http://suchanek.name/work/publications/kdd2006.pdf`.

Fabian M. Suchanek, Gjergji Kasneci, and Gerhard Weikum. YAGO: a core of semantic knowledge unifying WordNet and Wikipedia. In *WWW '07: Proceedings of the 16th International World Wide Web Conference, Banff, Canada*, pages 697–706, 2007. ISBN 978-1-59593-654-7. doi: 10.1145/1242572.1242667.

James Surowiecki. *The Wisdom of Crowds: Why the Many Are Smarter Than the Few and How Collective Wisdom Shapes Business, Economies, Societies and Nations*. Doubleday, May 2004. ISBN 0385503865.

D.M. Titterington, A.F.M. Smith, and U.E. Makov. *Statistical Analysis of Finite Mixture Distributions*. Wiley, New York, 1985.

Kosuke Tokunaga, Jun'ichi Kazama, and Kentaro Torisawa. Automatic discovery of attribute words from web documents. In Robert Dale, Kam-Fai Wong, Jian Su, and Oi Yee Kwong, editors, *IJCNLP*, volume 3651 of *Lecture Notes in Computer Science*, pages 106–118. Springer, 2005. ISBN 3-540-29172-5. URL `http://dblp.uni-trier.de/db/conf/ijcnlp/ijcnlp2005.html#TokunagaKT05`.

L. Vlaming, J. Smit, and T. Isenberg. Presenting using two-handed interaction in open space. In *Horizontal Interactive Human Computer Systems, 2008. TABLETOP 2008. 3rd IEEE International Workshop on*, pages 29–32, 2008. doi: 10.1109/TABLETOP.2008.4660180.

Max Völkel, Markus Krötzsch, Denny Vrandecic, Heiko Haller, and Rudi Studer. Semantic wikipedia. In *WWW '06: Proceedings of the 15th international conference on World Wide Web*, pages 585–594, New York, NY, USA, 2006. ACM. ISBN 1-59593-323-9. doi: 10.1145/1135777.1135863. URL `http://portal.acm.org/citation.cfm?id=1135863`.

Sholom M. Weiss and Casimir A. Kulikowski. *Computer Systems That Learn: Classification and Prediction Methods from Statistics, Neural Nets, Machine Learning and Expert Systems*. Morgan Kaufmann, 1990. ISBN 1-55860-065-5.

Lauren Wood, Arnaud Le Hors, Vidur Apparao, Steve Byrne, Mike Champion, Scott Isaacs, Ian Jacobs, Gavin Nicol, Jonathan Robie, Robert Sutor, and Chris Wilson. Document object model (dom) level 1 specification (second edition) version 1.0, 2000. URL `http://www.w3.org/TR/2000/WD-DOM-Level-1-20000929/`.