Social networks are very rich information sources that combined with data mining techniques can represent a real opportunity to extract user knowledge and improve decision making systems. Even if information about one user on a social network is not complete, most of social network's users are active on several social networks at the same time, so the combination of data coming from different social networks can provide a quite comprehensive description of the user and its context. This document presents a method which combines Data Mining and Semantic Web technologies, to extract information from several social networks, including semantic links between them. The collected information is then stored in a semantic format using an ontology (called the Context Ontology). The first goal of this work is to determine the available information, and design the suitable ontology for this context. At last the project describes a possible representation of the stored information.
Summary

1 Introduction ............................................................................................................................................... 3

2 Related Works......................................................................................................................................... 4
  2.1 Data mining on Social networks ........................................................................................................ 4
  2.2 The Semantic web .................................................................................................................................. 4

3 Ontology Conception................................................................................................................................. 5
  3.1 Data Mining ........................................................................................................................................... 5
      3.1.1 Facebook .......................................................................................................................................... 5
      3.1.2 Foursquare ....................................................................................................................................... 6
      3.1.3 Twitter ............................................................................................................................................. 6
  3.2 Context Ontology Modelling................................................................................................................... 6
      3.2.1 Person .............................................................................................................................................. 6
      3.2.2 Event .............................................................................................................................................. 7
      3.2.3 Friend ........................................................................................................................................... 7
      3.2.4 Geolocation ..................................................................................................................................... 7
      3.2.5 Organization ................................................................................................................................... 7
      3.2.6 Place ............................................................................................................................................. 7
      3.2.7 Photo ............................................................................................................................................ 8
      3.2.8 Video ........................................................................................................................................... 8

4 Proof-of-Concept ...................................................................................................................................... 9
  4.1 Social Network Crawlers ....................................................................................................................... 9
  4.2 Data recording ...................................................................................................................................... 9
  4.3 Visualizing data ..................................................................................................................................... 10

5 Evaluation and Discussions ..................................................................................................................... 11
  5.1 Extraction .......................................................................................................................................... 11
  5.2 Visualization ...................................................................................................................................... 11

6 References .............................................................................................................................................. 12
Context Ontology for Social Media

1 Introduction

Especially since the beginning of Web2.0, with the navigation’s speed it brought, social networks became a very important source of information. Most social networks are specific and can be a potential way to get information about different aspects of a single user. While LinkedIn contains data about his professional life, Foursquare is dedicated to his venues and Facebook can bring more data about his interests. This way, social networks are currently a real challenge within the ground of data mining.

Semantic web aims to bring semantic content to data only readable for a human, allowing the computer to be able to understand the context of these data. It includes creating objects and semantic links between these objects, under a form of graph understandable for a machine. This concept is called an ontology.

The existence of numerous social networks and their variety is such that it’s very frequent for a single person to be active on many websites, with different information on each profile. The technology brought with Semantic Web is a good way to link the different data available concerning a single user.

This document starts describing the importance of social networks as rich source of information for data mining. Then we present the information that can be retrieved from most used social networks, and how we designed the Context Ontology that is used to represent in a semantic format the collected information. The last section of the document presents the prototype that has been developed.
2 Related Works

This section presents some existing works that are the most relevant to our ontology design work.

2.1 Data mining on Social networks

Available information on social networks is various. There are two types of information that can be extracted.

On the one hand we consider data given directly by the user, on his profile. On the other hand, data can also be retrieved by his friends. For instance, analysing the position of the user inside the network and his links between people, using algorithms, can help to learn a lot of information concerning his social relations [1].

The direct extraction of data on networks is also more and more developed. The biggest social networks often provide API to allow the development of applications using their data. Numerous projects in relation with the extraction of data on Twitter or Facebook were developed (cf. [2] [3]).

2.2 The Semantic web

Semantic Web was initiated by the World Wide Web Consortium (W3C), and aims at the integration of semantic content within web pages [4]. So that the computer is able to understand and establish links between pages: present text in HTML pages is only readable for a human (NB: currently, search engine often uses the occurrences of specific words to make links between data, whereas semantic web would include content which would permit to directly identify two pages as talking about the same subject without deeper analyses [5]).

The core technologies upon which the Semantic Web is build are ontologies; they can be defined as the explicit specification of conceptualization of a domain [6]. They allow for a specific context to organize concepts under the form of a graph and this way present a model for a set of knowledge.

Several works, that use an ontology to link different information concerning a single person, already exist. The FOAF [7] project aims at combining different profiles of a user by using an RDF ontology. The analysis of the work of Ding et al. [8] brought an interesting view: the developers decided to use the email of the user as a stable link between different profile. The ontology was also designed taking inspiration on the project Schema.org [9] which provides small ontology examples concerning specific concepts like Person, Place, Organization, etc.

There are different possibilities, and so different languages, to link data with semantic web. Even if the syntax differs, it always consists in triples such as subject-predicate-object, the subject being the URI of the concerned to whom we associate a predicate type link to the object. The simplest provided syntax in Semantic Web is RDF/XML. OWL is the standard language to create ontologies. It allows to include restraints on the links between the stored objects. The designed ontology on this project only needs to include the notion of classes and sub-classes, but does not require the other constraints available with OWL. RDFs is a schema language which allow to define a structure for RDF data simpler, but also less powerful, than OWL (it can be considered an intermediate between RDF/XML and OWL). For querying, we use the most popular query language for RDF data: SparQL [10], a SQL-like query language for RDF data.
3 Ontology Conception

The aim of this work is to collect data on a user (what we call context information) and use this information in order to automatize and personalise the services provided by the public administration. The collected information is stored in a semantic format structured according to the Context Ontology.

This ontology was designed taking into account the data that can be extracted on the three main social networks used in the project (Facebook, Twitter and Foursquare). Nevertheless, we tried to build it to be generic while keeping it the most usable possible for this specific project. To guarantee this we have based our work on some existing ontologies (e.g. person ontology, place ontology) available on Schema.org. In particular, we have reused, extended and linked those ontologies to create our Context Ontology. The goal of this ontology is to facilitate (automatize and personalize) the execution of business process related to public services. This ontology is filled in with information extracted from the different social networks (it is worth to note, that we don’t collect all the information related to the user, but only the one relevant for the execution of the BPM).

3.1 Data Extraction and Mining

Once the main concept of the ontology where defined, the detailed conception of the model was done in an iterative way, by taking into account our needs from the business process point of view and the information that was available on each of the three networks. To do this, it was necessary to study the data access authorization given by the three websites. In particular, there are two possibilities to access data on these three social networks according to the provided APIs:

- The “visitor” mode: anyone can access a small subset of public data concerning the user
- The “connected” mode: we can access all the data concerning the user but it is necessary to have been granted the user permission

For our project in order to maximize the chance to find necessary information to fill in the ontology, it’s interesting to retrieve information about the user himself but also about his relations (friends, family, etc.). However, as mentioned before the APIs do not provide the same type of information concerning the “connected” user and the others (i.e. if a user asking for a service give the system the permission to access his data with the connected mode, the system will not have automatically the right to access in the connected mode also the data of the other persons related to the user (family, etc.)). It appeared very fast, that in fact, the information given on a non-connected user is very poor, and barely usable. To take into account this in our ontology we decide to focus principally on extracting the information from a connected user, and we only extract the id (and the name on Facebook and Foursquare) from the users who share a link with him: friends, author or persons tagged on a photo, video that belongs to the user or the one the user is himself tagged on, or the author of an event the user performs in.

3.1.1 Facebook

Facebook’s API provides a lot of various information about the user: their interests, favourites books, films, video games, music, events a user was into, organization he belonged to. It’s also possible to retrieve photos and videos, the person has posted, or has been tagged in. Facebook allows to see the organizations (work, school or club e.g.) and the events the user has been part of. For each photo, video, events and organization, we can also retrieve the id and name from the author, and the contributors (the tagged persons), information concerning the place and the date. We can also access the album of the user and the group he joined. At last it’s possible to retrieve a user’s friends (their id and name).
### 3.1.2 Foursquare

Foursquare social network is more focused on places than on persons, its API provides an interesting database concerning places, and less information about users. It’s possible to retrieve the user’s friends (id and name), basic information available also on the other two networks (which is listed later) and photos.

### 3.1.3 Twitter

Twitter information policy does not allow the access to many data, it mainly contains two types of data: basic information given by the person on his profile, and based on the tweets he sends. It’s possible to extract for example information concerning his place when he tweets (under the condition that the user has enabled to extract the geographical location of his tweets), or posted photos with the tweet.

### 3.2 Context Ontology Modelling

Based on the analysis of available information on the social networks presented above, we identified three types of data properties associated with a Person:

- Information considered as identical on the three networks: name (first and last name), home location, birth date, gender and relationship status.
- Information which can differ: email, bio (a small sentence the user writes to describe himself), picture profile, his ids on each network, photos, websites, phone and locations.
- Information on Facebook only: videos, event, information concerning the places he worked or studied at.

Based on the recoverable data and on retrieved ontology on Schema.org, we chose to define the different classes to model our data.

#### 3.2.1 Person

This is the most important class in our ontology, it represents a user, and every other entities are (directly or not) linked to a Person. Any information extracted about the user is stored as data property with this entity: name, home location, birth date, gender and relationship status are retrieved once in priority on Facebook. Bio, email, id, link and picture, which are retrieved on the three networks are stored each type under three sub-properties (e.g. the data property bio, has three sub-properties: facebookBio, foursquareBio and twitterBio). At last Person can be linked to the following data properties (only retrieved on Facebook): birthPlace, books, games, gender, interested in (the relationship he’s looking for, e.g. friends), interests, languages, movies, music, political, quote, relationshipStatus, religion.
3.2.2 Event

This entity represents the events a user went to (only available on Facebook). It’s linked to the following data properties: datePublished (the date when the event was published on Facebook), description, facebookId, name. They also can be linked with a Place, that describes its location, a Person (or a Friend) that is its author, or a Person (or a Friend) that is a contributor (a person tagged on the photo).

3.2.3 Friend

This entity represents the friends of a user, as said earlier we can often retrieve the id and the name of friends, so as long as we can’t have any additional information on a friend’s profile, we only keep him as a Friend, but if later we can retrieve his profile information, he’s stored as a Person. This class owns three sub-classes: one per social networks. It allows to differentiate Friend who only contains a maximum of two information (name and id), on one social network and Person.

3.2.4 Geolocation

It’s always linked to a Place and contains two data properties, latitude and longitude which describe geographic location of the Place.

3.2.5 Organization

It describes the work or educational organizations which the user belongs or belonged. When available, the following information are stored as data properties: beginning date, date the user joined the organization, category, club or school for example, facebookID, the organization existing only on Facebook, name, ending date, date the user stopped belonging to the organization, facebookLink, a link to the description page of the organization. They can be associated with a Place.

3.2.6 Place

It describes the locations, as you’ll see in Figure 1 it can concern the location where a user was found, or the location of photos, event, organization. We store as data properties its category, city or country o.g., city, the city where the Place is located, country, description, facebookLink, a link to a description page of the Place on Facebook, foursquare-Link, a link to a description page of the Place on
Context Ontology from Social Media

Foursquare, name, postalAdress, postalCode, state, twitterLink, a link to a description page of the Place on Twitter.

3.2.7 Photo

It describes a retrieved photo. The associated data properties are: datePublished, id, retrieved on facebook or foursquare, name. They also can be linked with a Place, that describes its location, a Person (or a Friend) that is its author, or a Person (or a Friend) that is one of the contributor (a person tagged on the photo).

3.2.8 Video

It describes a video. The associated data properties are: datePublished, description, facebookID, the event existing only on Facebook, facebookLink, a link to a Facebook page associated with the video, link, the URL of the video, name. They also can be linked with a Person (or a Friend) that is its author, and a Person (or a Friend) that is one of the contributor (a person tagged on the photo).
4 Proof-of-Concept

Based on the analysis of the three social networks presented before (Facebook, Foursquare and Twitter), we designed the Context Ontology used to store the information in a semantic format. This section presents the prototype we developed in order to collect, store and visualize the data.

4.1 Social Network Crawlers

The first handles the extraction of data on the three social networks, the extracted data are first stored in intermediate classes. This module contains one model per extracted object (photo, user, video, event, organization, location), and other functionalities which directly take care, on the first hand, of the authentication of the examined user, and on the other hand, of the extraction of his information (and the intermediate store within object). For the extraction of data on Twitter, we used Twitter4j [11]. It’s a useful library, which easily allows to authenticate a user, and retrieve the available data (directly on the profile or with his tweets) under the form of Java objects. Twitter is one of the firsts to propose this type of library, to access data. The library Scribe [12] was chosen to handle the authentication on Facebook, and Facebook4j [13] for the extraction of most data. For a few information concerning places (Facebook4j being not very complete for those above) we also used HTTP requests using Scribe. For the extraction of data on Foursquare, the library Scribe was used. To parse the JSON results, we used Jackson.

An application console handles the adding of a person to the ontology. To be able to add a person, we need to know each id associated with his profile, on the three social networks. Then following instructions, for each social networks, an authorization URL is generated. The user must copy this URL in his navigator. For Facebook and Foursquare, if he is not already connected to the network, the website will propose him to log, then he will be redirected on a page, else he will be directly redirected to this page. The URL of the page contains a code; this code must be then entered in console. For twitter it’s almost the same but the code is given only on the page.

4.2 Data recording

This module supports data storage. This includes: the transformation of each property of each object (obtained after extraction) into RDFs triples, adding these triples on the repositories, and every method to access these data for visualization. The storage of RDFs data is handled with Sesame, an open-source framework for querying and analysing RDF data [14].

To store the data into the ontology, the URI of each entity is composed with the following string character: concatenated with the id of the object. For objects whose type is Person, the Facebook id is chosen, if he doesn’t have any, we choose the Foursquare id, else the Twitter id. For the Place type objects, we directly take the name (replacing coma with dot and trimming white spaces) of the location instead of the id. For any other type of object, we take its id on the social network where it was extracted.

Before adding any object, we check if it’s already present on the repository, therefore the use of the id within its URI allows to avoid adding twice the very same object. To add the friends of the user, we first check if a Person exists within the database that owns the same id on the social network corresponding, if it does, we just create a link between these two Person entities. Otherwise, we create a new Friend object with the retrieved id (and eventually name).
4.3 Visualizing data

To give a visualization of the data stored in our repository, we decided to focus on the different places where a user was. The places are interesting because they are recoverable on the three social networks, moreover it is easy to have an interactive representation of this type of data, using a map. Sesame, the framework used for the storage, allows querying with many languages, such as SparQL.

This last module handles the information representation. This includes, querying for the needed information on the repositories, using SparQL requests. The represented locations are the places where a user was, the location of a photo or event where the user was tagged, or the location associated with an organization he belonged to (which the condition that these places are linked to a Geolocation property).

This module allows to generate a map in the browser, using the Google Maps api. The locations are associated with markers, which are placed on the map thanks to the latitude and longitude obtained with the Geolocation of the location. The figure 2 below is an example of a generated map.

The prototype proposes to choose a user, existing on the repository among a list and to view the locations which have been stored, associated with him. First a little window opens, with two components: the list of available users and a Load button. The user can choose the person to visualize, and then generate the map in his browser with Load.

![Figure 2 - Data Visualisation](image)

Each marker is associated with a location and an info window, which provides information. Green markers stand for location where the user took or was taken in photo. Blue markers stand for locations where the user went and sent tweet or changed his Facebook status. Red markers stand for any other type of marker. Clicking on any marker, an info window opens, this window gives complementary information linked to the location (when available):

- The type (photo, event, organization, birth place or home)
- The name of the location, the name of the event, photo, organization associated with the location
- The date associated with the location, and for an organization, the date he joined and quit it
- A link to a web page on the network where it was extracted containing more information
5 Discussions and conclusion

The developed prototype provides a way to link the information of a person extracted on three different profiles on social networks, and gives a visualization of these data, as it was the objective. The implementation of data extraction process and of data visualisation are discussed in this section along with suggestions for further improvements.

5.1 Data extraction

At the end of the project we have developed a functional prototype for the data extraction module and the crawlers for the three social networks, which can retrieve an interesting amount of information about the user. For what concern the data about the friends of the user the prototype can retrieve the id, and sometimes the name, of the friend. This id is necessary to create a new instance of the “person” class in the Context Ontology (ideally to add a new person it is important to have the three user’s ids on the three social networks, however if a person is connected to only one or two social networks, it’s possible to add the person to the repository, but if this person creates later a profile on the missing social networks, it is not possible anymore to update his information).

In the current version of the prototype information concerning user friends is retrieved using the “visitor” mode. To improve the quantity and quality of the information retrieved it could be interesting to add to the current prototype the possibility to ask the permission to the friend of the user to access his data with the “connected” mode.

During the developed of the prototype some problems were encountered to extract the information with the “connected” and “visitor” modes, however these problems were solved in the end.

To evaluate the proper functioning of the data extraction module, we have manually compared the data extracted and visualised with our prototype with the data available on the three social networks.

For what concern the “location” information we tested that each marked location on Facebook was also correctly marked on our representation. Moreover, we tested that the prototype is able to collected information from user’s friend network to complete the “context” of the user (we tested the system with one person that didn’t post any album or photos himself, and we checked that the ones that were displayed on the visualization interface were the pictures taken by other persons belonging to his network of friends.

Concerning the extraction on Twitter and Foursquare, the evaluation was more time consuming since we could not use the information visualized on the interface but we had to check manually by comparing if each location associated with a tweet was correctly displayed on the map. It is worth to note that Twitter API only allows to retrieve the last 300 tweets posted by the person, but for our tests this was not a problem. With the current version of Foursquare, it’s not easy to post photos associated with a location: it’s only allowed to add a new place in the database, and we estimated this was hardly doable for a simple test. So we could not test the extraction of locations on Foursquare.

5.2 Data visualization

The visualization module provides a graphical representation of the geotagged data of the Context Ontology (see Figure 3 – right). To test this visualisation we have compared it with the ones provided by Facebook (as shown in Figure 3 – left, Facebook also proposes a representation of the user’s locations on a map that displays the different location where the user was, based on his photos, status, event and work place). Although there can be different objects (photo, event) associated with the
same place, the prototype only displays the information of one object per location. That is due to Google Maps API who only allows one marker per location.

As further improvement of the visualisation module, it would be interesting to add an element showing the friends who were with the person at the same event. The information about the people who were with the user could be found directly on the social networks: knowing the tagged people on the user’s photo. This information could also be retrieved by deduction: analysing for example that two persons are close friends, and were both around the same place at this moment (even if they don’t appear both on the same photo), we could guess they were together.

![Figure 3 - Final Visualisation and comparison (Facebook visualization - on the left and our prototype visualization – on the right)](image)

Finally, for the visualization of places which are associated with photos, it would be nice to be able to show the photo itself when clicking on the marker. However, in the current implementation the retrieved URL associated with the photo, is a link to a Facebook page, that contains the photo.

6 References