KATHOLIEKE UNIVERSITEIT LEUVEN

TU Graz
Graz University of Technology

Master Thesis: Progress Report 1

Laurens De Vocht

Master in Computer Science Engineering
Master in de ingenieurswetenschappen: computerwetenschappen

Subject:

Scientific Profiling based on Semantic Analysis in Social Networks

Supervisors:

Dr. Martin Ebner
Prof. Dr. Erik Duval

Promotors:

Prof. Dr. Erik Duval
Prof. Dr. Nick Scerbackov

Academic year 2010 – 2011

Contents

Abstract

This the first report of a Master Thesis project in Computer Sciences at *Graz University of Technology* (TUGraz) and the *Katholieke Universiteit Leuven* (KULeuven). It is an overview on the first four weeks of research. First it discusses the problem statement, then the literature study is carried out, followed by a preliminary view on the software architecture. Finally an updated project plan is motivated.

Samenvatting

Dit is het eerste verslag dat kadert in een masterproef in de ingenieurswetenschappen: computerwetenschappen aan de *Katholieke Universiteit Leuven* (KULeuven) en de *Technische Universitat Graz* (TUGraz). We geven een overzicht van de eerste vier weken van het onderzoek en bespreken de probleemstelling. Vervolgens behandelen we de literatuurstudie. We geven een allereerste inzicht in de prille architectuur van het programma en hoe het een oplossing kan bieden. Ten slotte motiveren we de aangepaste versie van het projectplan.

List of Figures and Tables

List of Figures

List of Tables

Chapter

1 Introduction

This introduction gives a background overview for this Master Thesis as well as a definition of the problem. The purpose and the scope of this report are outlined.

1.1 Background

A major issue in the modern context of "Research 2.0" is the discovery and verification of other scientists (as twitter users and their tweets). Another trend is the linking of many unstructured data on the web.

In this thesis project a framework will be developed to analyze the microblogs of twitter users. The semantic analysis will be the basis for interlinking this data with the semantic web. It is very difficult to find out if someone really is of interest without having to read through dozens of blogposts. Proper interlinking should improve and speed-up the profiling process. Scientists will be able to learn how they are connected to others. Links can be built based on shared events or similar research interests.

There is a very interesting use case to illustrate this idea. When scientists are attending a conference, they might be interested in what is happening around them. Many attendees keep track of what's happening with their handheld or laptop. Especially he things they are blogging and tweeting about are of interest. They could discover new people attending the same seminar, since semantic interlinking connects them. For every tweet and user there is some kind of matching entity. The linking with the semantic web can supply and verify this identification and learns in which way users and their microblogs serve a certain research question, case, event or interest. This application should advice scientists or researchers and suggest connections with others.

1.2 Problem statement

The goal is to develop a scientific profiling application based on on existing developments, standards, libraries and community approved ontologies. The application will help to connect people that share similar interests.

SCIENTIFIC PROFILING Twitter feeds will be used as primary information source. The possibilities to structure and analyze unstructured data need to be investigated. They

must be summarized and linked to verifiable entities. The result should be an extended user-profile.
The semantic analysis is not going to use a new or customized ontology. An interlinking with several existing ontologies will be used instead. Finally an interface will be developed and designed to maximize the usability of the scientific profiling application that fits in the user‘s workflow.

USE CASE The implemented use case to test this application will be a contribution to the research field “Research 2.0”. The web based application should display an enhanced Twitter profile of a researcher. According to semantic analysis of user‘s tweets, research fields are carried out and connected with his working place (university) as well as participated conferences. The tool has to serve a real user‘s research needs. The actual relevance of found resources needs to be evaluated and observed. The interface must allow the user to browse the linked entities smoothly. The usability and relevance of produced results is much more important than the speed performance of the tool. The application is intended as a proof-of-concept and will point out the current state-of-the-art, research issues and limitations.

1.3 Purpose

This report is primarily intended for the supervisors and promotors of this Master thesis. Also everybody who is interested in the semantic web, microblogging and profiling might find some parts of this report relevant.
The next chapters discuss the literature study and software architecture. The literature study aims to gain more insight in the problem and more background information. This information leads to interesting insights concerning the architecture for a semantic profiling framework.

1.4 Scope

It is to be noted that neither the literature study nor the software architecture want to give a broad overview of the current semantic web and microblogging services. It is targeted as a carefully considered selection of articles that allows the development of the scientific profiling application. The architecture of the framework is being designed only with the problem statement in mind. At this time it is not part of the research to find out how this could be extended to other resources (besides Twitter) or targets (e.g. mobile applications). This report is limited to the research that has been carried out in the first four weeks of the project.

Chapter

2 Literature Study

An overview of the most important articles is given. The articles are presented in the following order: first articles handle the semantic web in general, then some cases of microblogging combined with semantics are discussed and finally this chapter presents a commented summary of some ideas that really support this specific case of scientific profiling in social networks.

2.1 A network of linked data

The semantic web represents a network of linked data. This data can be of any kind. It all started as a vision by World Wide Web guru *Tim Berners-Lee*. Since it was first introduced in 2001 the discussions have never stopped. There are those that claim it will disappear as slowly as it got popular, are against those, that ensure it will creep into all known-to-day web services. Ultimately the entire world wide web could form a huge semantic web. However interesting a study of the holistic view and the developments of its widespread reputation might be, it is not relevant at all for this project. It is more of interest to take a look at what is out there and which semantic web projects and tools can support the framework for the semantic profiling application.

2.1.1 Where it all started

Every study about the semantic web should include the very paper of Berners-Lee et al. published May 2001 in Scientific American[10]. In the article they presented the semantic web as a new form of web content meaningful to computers. They believed, and still do today, that it will unleash a revolution of new possibilities. The authors started with an example of the scheduling of an appointment by two busy persons. They both used the help of their software agents. Those agents were able to help them by being able to identify events, times and locations in their messages and link them to both their schedules. The authors called this concept: *the Semantic Web*.

The Semantic Web differs from the World Wide Web in the sense that it will bring structure to the meaningful content of Web pages, creating an environment where software agents roaming from page to page can readily carry out sophisticated tasks for users. According to the authors the Semantic Web is not a separate Web but an extension of

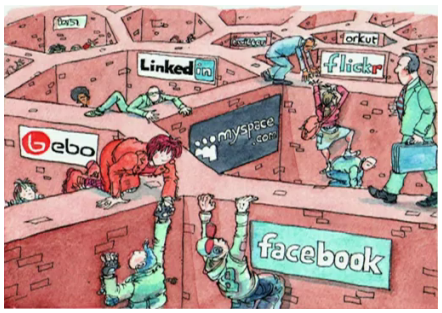

FIGURE 2.1: Walls between social networks as presented by Tim Berners-Lee.

the current one, in which information is given well-defined meaning, better enabling computers and people to work in cooperation. Like the Internet, the Semantic Web will be as decentralized as possible.

2.1.2 How the social web can be interlinked

Semantics in Twitter feeds and the profile of a user will be analyzed. An article by Bojars et al. "Interlinking the social web with semantics"[5] gives more insight in the relation between the current semantic and social web.

Bojars et al. discussed one of the most visible trends on the Web. Which is the emergence of Social Web sites, which help people create and gather knowledge by simplifying user contributions via blogs, tagging and folksonomies, wikis, podcasts, and online social networks. They noted that current online-community sites are isolated from one another (see Figure 2.1), like islands in a sea. The main reason for this lack of interoperation is that for the most part in the Social Web, common standards still do not exist for knowledge and information exchange. During the last couple of years, a lot of effort has gone into defining standards for data interchange and interoperation. The Semantic Web technology stack is well defined, enabling the creation of metadata and associated vocabularies. The Semantic Web effort is in an ideal position to make Social Web sites interoperable. Applying Semantic Web frameworks such as SIOC (Semantically Interlinked Online Communities)[3] and FOAF (Friend-Of-A-Friend)[2] to the Social Web can lead to a Social Semantic Web creating a network of interlinked and semantically rich knowledge.

2.1.3 Which layers the semantic web consists of

Reading a comment in the column "Trends and Controversies" in the magazine "IEEE Intelligent Systems" by Steffen Staab[18] led to an interesting paper by Peter Mika[18]. It supports the conviction that the integration of social network data from different sources is very important. The information produced in social networks has true value since it contains an extensive amount of knowledge. This knowledge is being communicated between people that are a members from a specific research group or community.

There are however some issues to be considered. Two in particular stick out from the thick proceedings volumes: ontology learning and ontology mapping. Ontology learning

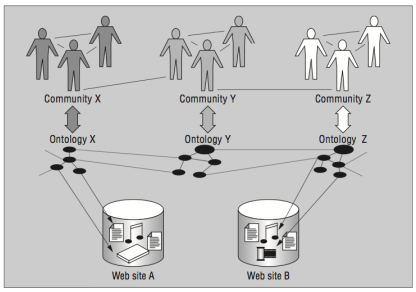

FIGURE 2.2: Three layers of the Semantic Web by Peter Mika

or extraction is the attempt to recreate a conceptual model from existing knowledge sources, in particular natural text. Ontology mapping (also known as merging, alignment, and so on) refers to finding and reconciling the relations between two or more conceptual models and creating a single model that captures their intentions and the relationships between them. They are explained very clearly in this article.

Staab stated:

> Social networks have interesting properties. They influence our lives enormously without us being aware of the implications they raise: How does a kind of fashion become en vogue? How does a virus spread and infect people? How does a research topic become a hot topic? Why are some companies successful and others are not? All these questions affect us, and understanding them by building and investigating computational models might give us a powerful tool to improve our health system, increase individual and general wealth, or just increase awareness about how the people around us actually influence our opinions, which we frequently believe that we shape.

Peter Mika considered a particular form of influence: the way that people agree on terminology and the phenomenon‘s implications for the way we build ontologies and the Semantic Web. In a nutshell, he reasoned that the Semantic Web will either include social networks‘ influence in its architecture or wither away.
The change of conceptualizations as communities evolve poses another challenge. This challenge is of course the “Ontology Mapping” he referred to earlier in his article. The more unstable knowledge is, the more difficulty we can expect in formalizing and sharing it on a large scale. Mika included an illustration in Figure 2.2 that shows how communities, ontologies, and content make up the three layers of the Semantic Web.

2.1.4 What semantic profiling is about

An interesting document[11] in which Dave McComb, President of “Semantic Arts”, explained out of his experience how one could conceptualize semantic profiling. He stated:

> Semantic profiling is a technique using semantic-based tools and ontologies in order to gain a deeper understanding of the information being stored

> and manipulated in an existing system. This approach leads to a more systematic and rigorous approach to the problem and creates a result that can be correlated with profiling efforts in other applications.

There is no better way to express this concept. If applied to the scientific profiling project: The semantic analysis of Twitter users‘ profiles should help in a deeper understanding of their scientific relevance. It will also create more opportunities to correlate “Research 2.0” applications.

2.2 Social networks in this decade

In the past few years the impact of social networks kept increasing. Because of the significance a study of several social networks‘ properties is useful. A number of articles highlight some specific properties that are of interest to this project.

2.2.1 Where the object centered sociality went

A five year old blogpost by Jyri Engestrom[8], co-founder of Jaiku, reads as if the problem is still actual. Engestrom notes that in the present social networks a very important part is often left out. It is the part that describes what connects people. Whether it is another person, a job, an event or a common interest. Many social networks make it difficult to disconnect from someone that is not known anymore or has an unknown origin. If social networks would become object centered Ð like they are in real life, then one would not have to deal with this issue. Online social connections would simply be build around the objects that connect people.

2.2.2 How online communities can be interlinked

In an article[6] Breslin et al. presented different types of online communities and tools that were at that time used to build and support online communities. Those communities are islands that are not interlinked. The authors presented the SIOC ontology. The goal of SIOC is to interconnect these online communities.

In the first section they presented the SIOC ontology. The ontology consists of two major parts: first, it contains classes and properties that describe discussion forums and posts in online community sites. Second, it includes mappings that relate SIOC to existing vocabularies such as FOAF and RSS. Breslin et al. elaborated on how the exchange, both importing and exporting data, can be executed. The core use of SIOC will be in the exchange of instance data between sites. Wrappers will allow to export instances of community site concepts such as forums or posts in RDF format. They can also allow to import SIOC instances to other non-SIOC systems. In the final section Breslin et al. talked about using SIOC Data. Given the ontology, the mappings, and the wrappers, they were now able to pose queries and add data to individual SIOC sites. They highlighted three aspects: browsing, querying and locating related information. The authors concluded that to tackle the challenge of adoption they have provided an upgrade path that allows a

FIGURE 2.3: A triple by Peter Morville

gradual migration from existing systems to semantically-enabled sites. For combination with other ontologies they have presented mapping to and from SIOC.

2.3 A story told in triples

A triple is a structure that connects a subject node with an object node by a predicate link, see Figure 2.3. Data generated in social networks can not easily be converted into triples. Those triples have then to be made available to other users. Ongoing research points out several of these challenges and issues. A few important are outlined in this section.

2.3.1 How semantic microblogging with Twitter could work

This project's framework will have to deal with short messages of less than 140 characters. This is called microblogging. Joshua Shinavier wrote a summarizing paper[15] on how this can be achieved. He introduced a semantic data aggregator which brings together a collection of compact formats for structured microblog content with Semantic Web vocabularies and best practices in order to augment the Semantic Web with real-time, user-driven data. Obviously this is the direction for the research in this project.

Shinavier's paper takes the approach of harvesting semantic data embedded in the content of microblog posts or of doing for microblogs what microformats do for Web pages. This is complementary to "Semantic wikis" and the "Microformats" community who aim to bridge this gap by enabling users to add small amounts of semantic data to their content. A number of compact formats have been proposed to allow users to express structured content or issue service-specific commands in microblog posts. So-called triple tags even allow the expression of something like a RDF triple. Microformats are subject to a tradeoff between simplicity and expressivity which heavily impacts community uptake. Shinavier gave the example of Twitter Data, Micro Turtle, Smesher and Twitlogic.

2.3.2 What a semantic microblogging architecture should look like

"SMOB" (Semantic MicrOBlogging) is an interesting system, because its architecture is similar to the kind of architecture needed to realize the scientific profiling application. SMOB has been described in an article[13] about Microblogging by Passant et al. It also described the implementation of an initial prototype of this concept that provides ways to leverage microblogging with the Linked Data Web guidelines. At the time of writing microblogging services were (and still are today) centralised and confined. Efforts are still to be made to let microblogging be part of the Social Semantic Web.

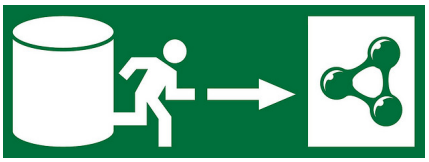

FIGURE 2.4: Exit to the semantic web.

The authors introduced classical microblogging and some of the issues it raises. The authors saw how the Semantic Web can help in getting rid of these issues and what it can offer that traditional services could not achieve. Passant et al. then gave an overview of microblogging and described why we should consider it and highlighted current issues. In the article they stated that they believe that the Semantic Web is an elegant solution to opening these data from proprietary data-silos. It is a solution to providing machine-processable data and metadata to microblogging as well as to delivering an open and distributed environment for microblogging.
They wrote about the architecture of a semantic microblogging service. In order to model the metadata of a microblogging service, they relied on two widely used ontologies on the Social Semantic Web: FOAF and SIOC.

To summarize this paper: it introduced the architecture and a first implementation of a distributed semantic microblogging platform. While existing approaches to convert microblogging services to RDF already exist for Twitter, their approach relies on a complete open and distributed view, using some standards of the Social Semantic Web. Moreover, some parts of their work, as the hash tag processing could be adopted to services such as Twitter to enable some semantics in existing tools.

2.3.3 Another case of data transformation

"SCOVO" (Statistical COre VOcabulary) is a vocabulary that supports systems where statistical data is being processed and linked to the semantic web. In the paper of Hausenblas et al. [9] this process and the use of SCOVO was explained. Their workflow is similar to the one being implemented in this project.

There are three important steps and every step has its specific tools that aid in the implementation. RDFication: with the help of domain vocabulary build RDF triples of the original data. Interlinking: this step results in linked data sets. Publication: here URI's are published of the RDF and (X)HTML over HTTP. The metadata can be deployed as SPARQL endpoints + RDF Dumps, RDF XML or XHTML + RDFa.

The authors compared this approach with two others: D2R Eurostat and 2000 U.S. Census in an overview table. It is important to note that all approaches have their limitations. One can select an approach depending on what dataset is being dealt with and what target system is involved.

2.3.4 How mining microblogs using semantic technologies can be done

The framework for the semantic profiling tool fits like a puzzle piece in a bigger system that is being developed in the research group "Social Networked Learning" at *Graz University of Technology*. Selver Softic of Infonova GmbH and Ebner et al. of the "Social Learning" department at TUGraz recently wrote a paper[16] about their ongoing research efforts aiming at knowledge discovery. They are aiming to provide a scientific architecture paradigm for building semantic applications that rely on social data.

For example they worked out an approach for interlinking and RDFising social e-Learning Web 2.0 platforms like ELGG based on semantic tagging and Linked Data principles[17]. A special module called "SID" (Semantically Interlinked Data) was developed to allow existing tagged and published user generated content an easy entrance into the Web of Data and to enrich it semantically on the other hand.

At the moment Softic et al. are focussing on data from Twitter. For this purpose they have implemented a tool "Grabeteer"[12] for storing and caching social data. In this paper they outlined the architecture for a system that can extract, structure and link the data grabbed from Twitter by the Grabeteer. They introduced the interesting aspects about microblogs, how far they correspond with ideas from other research areas like Semantic Web or Linked Data. They also tried to answer how far those two areas can be combined to gain more knowledge and mine usable data out of social context of microblogs. Finally they presented an architectural paradigm approach that delivers the answer to specified research issue. This architectural paradigm is the basis for the software architecture described in chapter 3.

2.3.5 Semantic Web Pipes for Semantic Mash-Ups

Something very promising is the concept of "SWP" (Semantic Web Pipes) similar to "Yahoo Pipes". At the DERI institute Le-Phuoc et al. have developed and tested a SWP system: "DERI Pipes"[7]. They presented the pipe concept[14] as a good basis for semantic web applications using RDF. The authors said that the use of RDF data published on the Web for applications is still a cumbersome and resource-intensive task due to the limited software support and the lack of standard programming paradigms to deal with everyday problems such as combination of RDF data from different sources, object identifier consolidation, ontology alignment and mediation, or plain querying and filtering tasks. Architectural styles have been around for several decades and have been the subject of intensive research in other domains such as software engineering and databases. They based their work on the classical pipe abstraction and extend it to meet the requirements of Semantic Web applications using RDF.

Le-Phuoc et al. found that the existence of standards and defacto standards for publishing RDF, key problem in systems processing RDF are:

> The data is fragmented; may be incomplete, incorrect or contradicting; partly follows ontologies, often with ontologies used wrongly or inconsistently, to name a few, and thus needs to be "sanitized" before it can be processed. A

specifically cumbersome problem is the use of different identifiers denoting the same object which need to be unified.

Web pipes are “live”: they are computed on demand when requested via an HTTP invocation, and thus reflect an up-to-date state of the system (which can be detrimental as well in some scenarios where caching would be applicable). The authors then continued with an example to motivate the use of semantic web pipes and give a concise overview how it works. They sketched the main functionalities and gave an overview of all the important operators. They also discussed the system design and implementation of their version of SWP. Finally they evaluated the system by means of a case study. The authors discussed some general remarks about the performance issues and commented on the evaluation methodology (cognitive dimensions of notations). This is an interesting concept that could greatly support the semantic profiling framework. At the time of possible use, in a later development phase, they should be investigated in more detail.

2.4 Conclusion

This chapter focused on some aspects of the semantic and the social web. The semantic web was presented as a network of linked data. Some challenges about how the social web can be interlinked were outlined. Finally ongoing research projects showed that it is possible to translate social web data into triples. But the result of this process is still not accessible to casual users and the information has to be linked more accurately to ontologies to create more relevant RDF data sources.

Chapter

3 Software Architecture

The terminology about what is being developed is not yet strictly defined. This is because at this point it is not sure if the semantic profiling framework, that is under development in the first part of this project, will be implemented as a web service or rather as a distributable package. Both can be used to support the user interface that will be developed in the second part of the project.

3.1 Design specifications

The framework has to support at least the scientific profiling application that meets the requirements to the use case presented in the chapter 1. Agile development suggests to work to use cases. Features will be added and implemented only if they are needed in a use case. The implemented features for the framework will be limited appropriately.

Based on the research work at TU Graz [17] the design consists of three layers: a data extraction layer, an interlinking layer and an analysis layer. In addition a programming interface to this framework must be provided. At this point the main focus of the research is on the specification of the extraction layer. This is marked green in the diagram Figure 3.1.

The extraction layer is modeled as a bottom-up only system. This is because there is no real interaction with the above layers. The only request it has to handle is: "give me all data about a person". The other layers will be looked into as soon as the first layer is being implemented. Before this layer is finished, the development of the next layer must start and so on. An iterative development plan supports this method. It is explained in chapter 4.

3.2 Extraction layer

The extraction layer collects data from a person from Twitter and the Grabeteer. This data is collected in a set of classes. These sets are categorized in two models: the “user microblogs model” and the “user profile model”. The user microblog model gathers all data from the tweets it gets from Grabeteer. This data will be requested directly from the database using MySQL queries. The user profile model parses the user profile with help of the Twitter API. These models serve a class that annotates the data using relevant entities

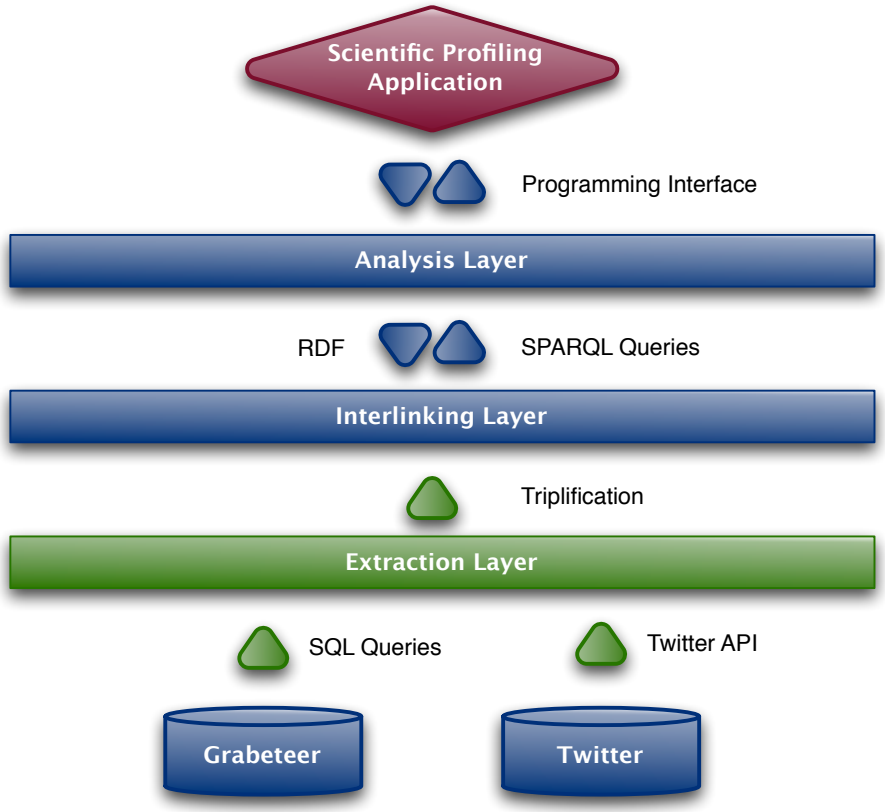

FIGURE 3.1: The semantic profiling framework design.

from ontologies. This annotated data will then be tripliflied in another class. The result of the extraction is a collection of annotated data in the form of triples. These triples are sent to the interlinking layer. Figure 3.2 illustrates this concept.

3.3 Other layers

This section is a vague concept. It is subject to change in the upcoming iterations. Nonetheless it is good to have an idea of what will happen with the data after the extraction and triplification.

INTERLINKING The interlinking layer will store the triples created in the extraction layer. They will be used as SPARQL endpoints to other ontologies such as DBPedia and GeoNames.

ANALYSIS The analysis layer makes an abstraction of the underlying RDF system. It provides an easy access to the underlying layers. The most important function is that it can translate high level information requests to SPARQL queries. It is actually an analysis, since it tries to combine and match a certain information need. It will not just dump the RDF data into another format.

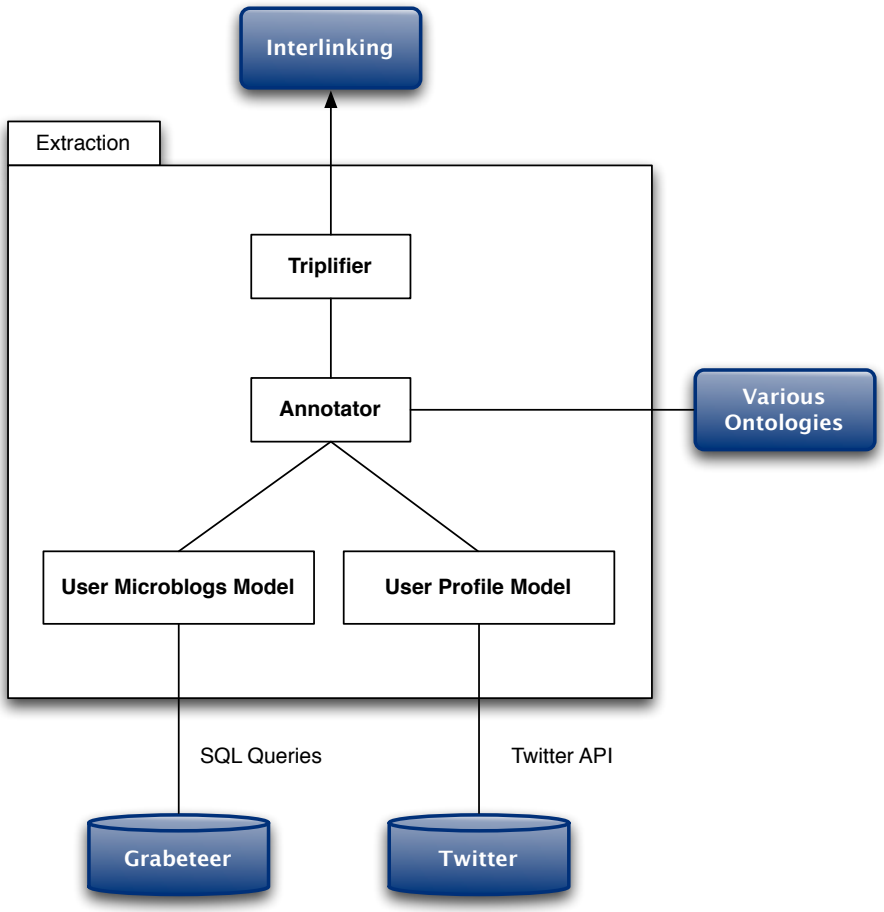

FIGURE 3.2: The extraction layer represented as a package.

3.4 Implementation considerations

Some early tests gave a sneak preview into the complexity of how to implement several of the aspects of the framework. For now it is sufficient to consider the ontologies that are used to annotate the data. Also the kind of triple store to use is being looked into. The triple store will have a central role in the interlinking layer.

ONTOLOGIES Research in the literature study in chapter 2 made clear: the following ontologies will be used in the extraction layer. The Dublin Core [1] will help to unambiguously describe the metadata. The FOAF project [2] and SIOC project [3] make it possible to represent the Twitter users and their online activities. The activities of Twitter users are mostly contained in their microblogs.

TRIPLE STORES Finding a suitable triple store requires a very resource intensive comparative study. This does not guarantee a solution. It might even lead to more confusion. There is no point in browsing through all available triple stores to determine the best one. They all are very different, use different API's and storage backends. Some have a native

store and others use a RDMS. There are benchmarking tools [20] that could form a basis for such a study. “BioPortal” performed some of these benchmarks on a selection of triple stores. Unfortunately they left out their conclusion in the public version of their report [4]. The W3C tested how well SPARQL is covered by most of the stores [21]. The W3C also listed triple stores [19] that do their job quite fast and reliably. So the only troubling factor left is the adoptability, how easily it can be integrated. The final decision will have to be made at the start of the development of the interlinking layer.

3.5 Conclusion

The software architecture described in this chapter will grow together with the project. An iterative development schedule makes this possible. The framework that will support the scientific profiling application is organized in three layers. An extraction layer, an interlinking layer and an analysis layer. The extraction layer will collect data from a user from Twitter and the tweets from Grabeteer. This data will be transformed into triples and interlinked with various ontologies and represented as SPARQL endpoints in the interlinking layer. The analysis layer is still not defined. But it will do more dan just transferring triples from the triple store to the application.

Chapter

4 Project Plan

Now the project plan is carried out. The system is shortly described. This chapter explains in more detail the work that has been done so far. It shines a light one the next few iterations. Finally the schedule summarizes this entire chapter

4.1 Overview

The choice for a plan with an iterative development allows agility. This ensures that every cycle evaluates the previous one and builds up to the next one. If adjustments have to be made they will be scheduled for the next iteration. Sufficient margin guarantees that all important milestones, will be met. The important milestones are at then end of January, March and May.

Because writing an effective report requires a lot of dedication, they are not included in the iterative development system. The concrete details about how this is implemented, are in the schedule section of this chapter.

4.2 Previous iterations

So far two iterations of the research have been carried out. A quick description follows in this section.

4.2.1 Iteration 1

A selection of literature informed about the current state of research. This selection serves as the basis repository for the next few months. To make it easy accessible, the entire library has been put online on a Mendeley account. Some particular papers turned out to be very interesting as a starting point for this project. They were studied more in depth in the second iteration.

4.2.2 Iteration 2

The previous iteration identified some interesting papers. They formed the basis of articles for the literature study in this thesis project. The summaries and comments are described

in the dedicated chapter 2. Furthermore a blog was setup to keep track of the research and development efforts made. Some early tests on existing systems that could support the semantic profiling framework were performed.

4.3 Upcoming iterations

The first layers, which are very low level, will be developed. The results from the tests at the end of iteration two will serve as a starting point. Also some tools and frameworks are already excluded, since the early evaluation proved them not suitable for this case.

4.4 Schedule

The following schedule represents the project plan and how it is being carried out. Details are in table 4.1.

4.4.1 Notes

During the first part of the plan the semantic and the social web is researched. This is the basis for the development of a semantic profiling framework and API. This API will be the foundation for the development of an application that fits the Research 2.0 use case introduced in this project. It is worth noting that the second part foresees more time to perform all tasks. This is necessary as at the end of that part the final thesis report must be written.

4.4.2 Changes

In the previous schedule there was no dedicated time to write reports. As this report took much more than the initially foreseen 8 hours to prepare, a time slot for each of the reports is appointed. The terminology in the schedule has been adapted with the appropriated names from the updated architecture description. Furthermore the workload of 24 hours (3 full days) a week turned out to be not achievable. The workload is now adjusted to 20 hours per week (2 days and a half), this is more realistic.

4.5 Conclusion

Not many changes were necessary to the original project plan. The schedule is updated with dedicated time for writing reports. The descriptions of the tasks in part 1 are now more detailed. The enitre schedule from the project plan is expressed in a table in this chapter.

	From	To	Weeks (#)	Work Load (est. hours)	Target/Task
PART 1 TUGraz	**4-Oct-10**	**24-Jan-11**		**Main objective**	**Framework development for Semantic analysis of twitter feeds and extended user profile synthesis**
Iteration 1	4-Oct-10	17-Oct-10	2	40	Get familiar with current research (papers)
Iteration 2	18-Oct-10	30-Oct-10	2	40	Research and evaluate relevant aspects more in depth
Report 1	1-Nov-10	7-Nov-10	1	20	Write first report
Milestone 1		**8-Nov-10**	**1st written report**		
Iteration 3	8-Nov-10	21-Nov-10	2	40	Develop Extraction Layer
Iteration 4	22-Nov-10	5-Dec-10	2	40	Test Exteraction Layer, Develop Interlinking Layer
Report 2	6-Dec-10	12-Dec-10	1	20	Write second report, prepare first presentation
Milestone 2		**13-Dec-10**	**2nd written report & 1st presentation**		
Iteration 5 (Christmas)	*14-Dec-10*	*2-Jan-11*	*2 (+ some Holidays)*	*40*	*Margin: used for unfinished work in it4 - start it6 earlier*
Iteration 6	3-Jan-11	16-Jan-11	2	40	Test Interlinking Layer, Develop Analysis Layer / API
Report 3	17-Jan-11	23-Jan-11	1	20	Write report and integrate with the 1st&2nd reports
Milestone 3		**24-Jan-11**	**End of work at TUGraz report**		
TOTAL PART 1			**15**	300	
PART 2 KULeuven	**25-Jan-11**	**30-Jun-11**		**Main objective**	**Develop a user interface that fits in a scientist's 'Research 2.0' workflow**
Iteration 7	25-Jan-11	7-Feb-11	2	40	Find out more about Research 2.0 applications & challenges
Iteration 8	14-Feb-11	27-Feb-11	2	40	In several iterations try to develop a solid user interface and implement it in an appropriate technology. Try optimize integration capabilities of the framework/API developed in part 1. Gather real user feedback! Evaluate the usability of the semantic analysis and profiling with this interface.
Iteration 9	28-Feb-11	10-Mar-11	1,5	30	
Iteration 10	11-Mar-11	20-Mar-11	1,5	30	
Report 4	21-Mar-11	27-Mar-11	1	20	
MIlestone 4		**28-Mar-11**	**3rd written report**		
Iteration 11	29-Mar-11	11-Apr-11	2	40	
MIlestone 5		**12-Apr-11**	**Second presentation**	8	
Iteration 12 (Easter)	*13-Apr-11*	*1-May-11*	*2 (+ some Holidays)*	*40*	*Margin*
Iteration 13	2-May-11	15-May-11	2	40	Optimize implementation of the system.
Report 5	16-May-11	29-May-11	2	40	Write final report
Milestone 6		**30-May-11**	**Final written report**	8	Review final report
Report 6	30-May-11	12-Jun-11	2	20	Preparation for final presentation
MIlestone 7		**End of june**	**Final presentation**	8	Review final presentation
TOTAL PART 2			**16**	364	
TOTAL			**31**	**664**	
			Avg work load	*21*	
			Margin	*80*	

TABLE 4.1: The research schedule.

Chapter

5 Conclusion

The literature study in chapter 2 highlighted some issues and challenges in the current semantic web. It shows that to make the social web a fruitful source for data there is still a huge leap forward needed. Both accessing and connecting the data are important issues. Social networks are like isolated islands. The information contained in there is just simply viewed by a few people and then stored. After storage it is not put into further practical use.

The architecture of the framework consists out of three layers: a data extraction layer, an interlinking layer and an analysis layer. An API, either a web service or a distributable package, will provide high level support for a scientific profiling application. The design will grow more specific as the project evolves. An iterative development system will make this possible.

The project plan foresees several iterations. This allows agility in the development. In every iteration the previous one is evaluated. If changes are necessary they will be scheduled for the upcoming iteration. This process will continue cyclically till a major milestone is reached. There is enough margin to ensure that the major milestones can be met.

Bibliography

[1] Dublin core meta data initiative. URL: http://dublincore.org/.

[2] The friend of a friend (foaf) project. URL: http://www.foaf-project.org/.

[3] Semantically interlinked online communities project. URL: http://sioc-project.org//.

[4] BioPortal. Comparison of triple stores. URL: http://www.bioontology.org/wiki/images/6/6a/Triple_Stores.pdf.

[5] U. Bojars, J. G. Breslin, V. Peristeras, G. Tummarello, and S. Decker. Interlinking the social web with semantics. pages 1–12, May 2008.

[6] J. G. Breslin, A. Harth, U. Bojars, and S. Decker. Towards semantically-interlinked online communities. *The Semantic Web: Research and Applications*, pages 500–514, 2005.

[7] DERI. Deri pipes. URL: http://pipes.deri.org.

[8] J. Engestrom. Why some social network services work and others do not - or: the case for object-centered sociality. URL: http://www.zengestrom.com/blog/2005/04/.

[9] M. Hausenblas, W. Halb, Y. Raimond, L. Feigenbaum, and D. Ayers. Scovo: Using statistics on the web of data. *The Semantic Web: Research and Applications*, pages 708–722, 2009.

[10] T. Lee, J. Hendler, and O. Lassila. . . . The semantic web. *Scientific American*, Jan 2001.

[11] D. McComb. Semantic profiling - an approach to understanding data in an existing system. URL: http://semanticarts.com/articles/semantics-and-ontologies/semantic-profiling, Sep 2004.

[12] H. Muhlburger, M. Ebner, and B. Taraghi. @twitter try out #grabeeter to export, archive and search your tweets. pages 1–9, Aug 2010.

[13] A. Passant, T. Hastrup, U. Bojars, and J. G. Breslin. Microblogging: A semantic and distributed approach. *Proceedings of the 4th Workshop on Scripting for the Semantic Web*, 2008.

[14] D. L. Phuoc, A. Polleres, M. Hauswirth, G. Tummarello, and C. Morbidoni. Rapid prototyping of semantic mash-ups through semantic web pipes. *Proceedings of the 18th international conference on World wide web*, pages 581–590, 2009.

[15] J. Shinavier. Real-time# semanticweb in<= 140 chars. *Proceedings of the Third Workshop on Linked Data on the Web (LDOW2010) at WWW2010*, 2010.

[16] S. Softic, M. Ebner, H. Muhlburger, T. Altmann, and B. Taraghi. @twitter mining #microblogs using #semantic t echnologies. pages 1–12, Sep 2010.

[17] S. Softic, B. Taraghi, and W. Halb. Weaving social e-learning platforms into the web of linked data. pages 1–9, Jul 2009.

[18] S. Staab. Social networks applied. pages 1–14, Jan 2005.

[19] W3C. Large triple stores. URL: http://esw.w3.org/LargeTripleStores.

[20] W3C. Rdf store benchmarking. URL: http://esw.w3.org/RdfStoreBenchmarking.

[21] W3C. Sparql implementation coverage report. URL: http://www.w3.org/2001/sw/DataAccess/tests/implementations.