University of Milan Bicocca
Department of Computer Science and Communication Technologies
Master Degree in Computer Science

STAN: an incremental semantic annotation tool for tables on the web

Supervisor:
Dr. Matteo Palmonari

Co-supervisor:
Dr. Riccardo Porrini

Thesis presented by:
Davide Brando Preda
735850

Academic Year 2014-2015
25/11/2015

Tables are a way to express important information that are difficult to put in free text. The table structure communicate the content's semantic: columns group elements of the same type or domain, while rows group more attributes related to the same subject. For humans it is easy to understand this semantic from a table but it is not the same for a machine that needs to understand a web table.
Tables are one of the few practices for structuring web data and probably is one of the most used. Indeed in Google's indexed web documents there are more than 154 millions on relational tables. In this thesis we focus on web table semantics, a problem that is part of a wider context that is the understanding of data on the web. Most of the web information is processed by search engines as simple unstructured strings of text. So in search mode the key to decide if a document should be considered relevant is based on terms frequency and similarity metrics. Yet in this way the document's semantic is lost and means that a machine could not understand the content of a web page.
The Semantic Web tackles this kind of problems defining standards to structure data on the web. The web is always been based on HTML language that allows to structure a web page from a graphical point of view in order to display it in a browser. However HTML doesn't allow to structure data in order to let them be understandable by a machine. In the last years the use of semantic annotations has spread on the web. Semantic annotations combined to HTML language allow to give a precise semantic to web pages elements. Using annotation languages like RDFa it is possible to enrich pages contents with useful metadata. These metadata are then extracted by search engines and web crawlers to provide users with a more exhaustive and rich research experience. For example Google use those annotations to automatically build informative panels next to the usual search results.

Semantically annotating a web content means enrich it in order to assign a known meaning. In particular annotation aims to assign the semantic of a concept or property from an ontology to a specific content. An ontology is a formal representation, shared and explicit of the conceptualization of a domain of interest and can be interpreted by a machine. So exploiting annotations and ontologies knowledge a machine can understand a web page content or a table's data. Once the data are structured it is possible to integrate them. After integrating them it is possible to provide an answer for complex queries that involve the understanding of data from heterogeneous sources. For example, given two web tables about statistical data on earthquakes in the last century in Italy and data on the number of buildings constructed in the Italian cities, someone could be interested in finding a correlation between these events. To carry out this process is needed to understand the semantic of the two tables, annotate them and finally integrate them. The annotation problem is known in the Data Integration research area. In classic Data Integration indeed one of the integration techniques creates a virtual integrated dataset by defining mappings between local schemas and a global schema. So annotation can be seen just like the definition of a mapping towards a global schema that is represented by an ontology.

The aim of this thesis is to focus on the data in table format and to develop a tool that allows to perform semantic annotations on them. The result of this thesis is STAN, a web application available at `http://stan.disco.unimib.it/` which allows via a graphical interface to import a table, annotate it with semantics from an ontology and then export the results of the defined mapping (figure 1).

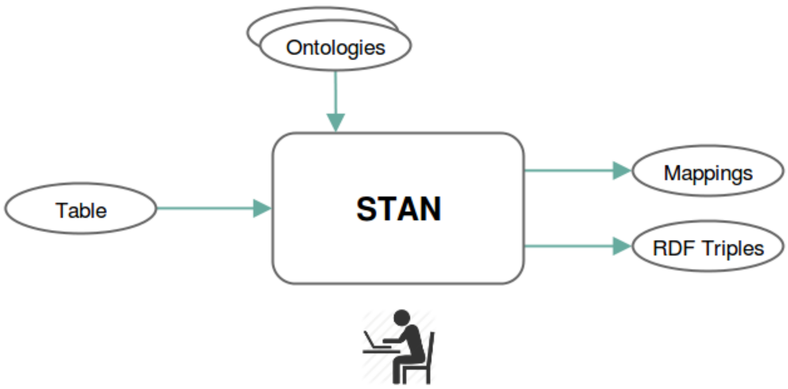

Figura 1: STAN's annotation process.

The development of the tool has been driven by two real case studies through which it has been possible to study and understand the issues related to the data integration. The case studies refers to different domains and they have given rise to different needs and problems. The first case study is linked to social and economic issues and deals about the decision for closing 47 schools in Chicago to reduce the operating costs. The second one concerns the world of eCommerce, in particular 7Pixel company, and deals about the problem to dynamically handle and integrate commercial offers from heterogeneous sources. The analysis of these two case studies allowed to deduct some of the problems and challenges that the implemented tool addresses. In particular, the major themes are: the difficulties in the integration of heterogeneous data and their semantic interpretation, the need for automation of the process of integration in case of large volumes of data and the need to make the process executable by people who do not have technical skills.

The main contributions of this thesis are:

1. The problem formalization and in particular the formalization of the possible annotation models.
2. The implementation of a tool in support of table annotation.
3. An experimentation that compares the annotation algorithm used in the tool and an existing one from state of the art.

Starting from the analysis of the state of the art and trying to annotate the real tables from the case studies, the existence of some recurring patterns showed up while annotating a table. These patterns were then formally defined in a number of possible annotation models. The models were then used with a dual purpose: to better understand how and which functionality develop in STAN and to make a comparison between the various existing annotation tools.

STAN implements only a subset of the defined models and in this lacks with respect to other annotation tool from the state of the art. However STAN has certain characteristics that differentiate it from other tools. First STAN allows a user to define incrementally its own ontology during the annotation process and this allows users to use their own semantics for structuring data. Another difference between STAN and other tools is the architecture designed purely for the web. Indeed STAN is a web application used online that allows each user to have his own personal workspace. The other tools instead are meant to be downloaded and used by a single user on his machine. STAN also is the only tool that uses an *instance based* algorithm for automatic annotation based on a knowledge base. Finally STAN is the only tool that exposes a public API service allowing third party applications to perform automatic annotation of groups of values. The API service was designed specifically thinking about the case study of 7Pixel. One aim of the thesis, in fact, is to make STAN a annotations provider within 7Pixel's systems and APIs are the simplest way to allow integration with 7Pixel's legacy systems.

The experimentation finally aims to evaluate the annotation algorithm in terms of quality and performance in relation to the state of the art. Experimentation has in fact been conducted to study the algorithm under two different points of view: to evaluate the quality of suggestions in the specific case of eCommerce and to evaluate the speed performances in relation to some variables. The results of the experimentation showed that on average the approach from state of the art provides suggestions of better quality compared to the algorithm implemented in STAN. However there is a huge difference in the quality of the suggestions given for the columns containing numeric values. For this type of columns, in fact, algorithm from state of the art gets quality results significantly lower than STAN especially when the volume of information available increases. The poor results obtained by STAN in textual columns annotation instead can be easily improved by introducing some domain specific heuristics. For example recognize certain recurrent patterns as "http://" help the correct identification of columns as *Link* and *Image* for which the algorithm currently gets very bad outcome.

The development of STAN has been designed mainly for two purposes: the integration in 7Pixel's systems and the spread as an Open Source annotation tool. At the moment these two souls coexist and the different application logic is only specified at deployment time. Probably in the future these two souls may diverge increasingly with the introduction of new features then handle the logic that differs only at deployment time might tend to become too onerous and complex. For this reason in the future it is expected to split STAN into two projects extolling individually its two cores. In the 7Pixel case we will focus on making the annotation algorithm more precise in the domain of eCommerce by introducing some specific heuristics which to provide better annotation suggestions. Outside 7Pixel instead we will attempt to promote STAN as an Open Source tool for the annotation of data in Open Data tabular format and for this reason we will focus on the graphic and on the adding of new features.