Searching Time Series via Semantic Annotations

Lucas Oliveira Batista[1], **Claudia Bauzer Medeiros**[1] **(advisor)**

[1] Programa de pós-graduação em Ciência da Computação do Instituto de Computação da UNICAMP – Universidade Estadual de Campinas
CEP 13083-852, Campinas - São Paulo - Brazil

lucas.batista@students.ic.unicamp.br, cmbm@ic.unicamp.br

Level: Master's Degree

Admission: March 2013

Qualifying exam: 09^{th} October 2013

Expected Conclusion: March 2015

Concluded stages: Credits; Qualifying exam; Definition of time series semantic annotation model; Implementation of first software prototype.

Future stages: Improve the software and evaluate the time series semantic annotation model; Develop algorithm to search time series via semantic annotation; Evaluate this algorithm; Write paper and dissertation; Master's degree defense;

Abstract. *Time series are used in several domains of knowledge. During their analysis, experts often create or analyze associations between time series and annotations. In order to study a problem, for example, patient behavior or crop patterns, experts need to search and correlate several time series. However, finding appropriate series related with a problem is a difficult task. Search is usually performed using a few parameters, such as series geographic location. Annotations may be used to help the search using string match. Given this scenario, this paper discusses a work in progress to design and partially develop a software framework to search time series via semantic annotations. It will support experts in the correlation of time series, foster collaboration among experts, and allow the use of Linked Data concepts to aggregate knowledge to content. This paper extends the short paper accepted for BRESCI - Brazilian Workshop e-Science 2014. The extensions include a time series semantic annotation model, implementation details, and a longer theoretical related work section.*

Keywords: *Search for time series, semantic annotation*

1. Introduction

Time series are used in several knowledge domains, such as economics (monthly unemployment rate), meteorology (daily temperature) or health (electrocardiogram). An efficient search for time series helps the task of analyzing these series, for instance, to perform forecasts, identify patterns, find the origins of certain phenomena, and so on. Experts often use series annotations to help this analysis.

An annotation establishes a relation between the annotated data and the annotation content [Oren et al. 2006]. Time series annotations are potentially made by different researchers or research groups and they are generally created in text format using annotation specific languages and stored in files. Due to the fact that there are many standards in a single domain, annotations have a large heterogeneity. It complicates the search, sharing and integration of data.

In several situations, scientists need to correlate many kinds of series to study a problem. The search for relevant series with respect to the problem is expensive and takes a long time. Conventional systems perform the search for time series using either time series similarity, parameters that are matched with time series values, or annotation textual match. However, these approaches may not be enough to find series related with a problem.

Suppose that a researcher needs to analyze the impact of corn production on local economy of a region. Several series are related to corn production, for instance, temperature, humidity, harvested amount, and so on. Searching for relevant series presents challenges, for instance, "corn" is cultivated in different parts of the world and, unless the series are georeferenced, the search by keyword "corn" will return too many series. Furthermore, each region has a different nomenclature for "corn" and many relevant series may not have this specific annotation.

Given this scenario, the goal of this work is designs and partially develops a framework to support series search via semantic annotations, thereby helping the correlation task. This paper extends the short paper accepted for BRESCI - Brazilian Workshop e-Science 2014. Extensions are: a time series semantic annotation model and implementation details (Section 5), and a longer related work section (Section 2).

2. Theoretical Basis

Annotations are data about data and they are used to provide extra information that may be relevant to data analysis [Silva 2013]. Annotations can have different information depending on the annotated data, and on who performs the annotation. Besides, annotations may be made in many formats, for example, video, audio or text. This work focuses in textual annotations of time series.

A *semantic annotation* is the description of a digital resource according to its semantics [Sousa 2010]. [Sousa 2010] defines semantic annotation as triples $< s, m, o >$, where s is the subject described, m is the label of a metadata field, and o is an ontology term that semantically describes this subject. We extend this notion to a more generic structure.

3. Related Work

The goal of this work is to provide the users with tools that help the search for time series, taking advantage of semantic annotations. Therefore, the literature review focuses on highlighting research on time series annotations (2.1) and time series search (2.2).

3.1. Time series annotation

There are many tools that support series annotations - for example, [Pressly 2008, Silva 2013]. Both enable associating multiple textual annotations to parts of a series. In the first, experts may just visualize annotations made by others. In the second, experts may collaborate modifying annotations made by others.

Such tools present a few disadvantages. First of all, they store annotations as free text, which hampers search, sharing, integration of data and their interpretation by machines. Second, they do not consider the ambiguity of annotation contents. These problems may be attacked using semantic annotations.

The Tatoo Framework [Pariente et al. 2011] enables the semantic annotation of environmental resources and has a time series semantic annotation model. However, this model is based on time series observations and has a different purpose. Tatoo performs semantic annotations to semantically represent the temporal data. Our work focus on annotations that are made by experts, attaching extra information to series.

3.2. Search for time series

According to [Gao and Wang 2009], the search for time series refers to finding, from a set of time series, the series that satisfy a given search criterion. This search may be done using a model or statistics of series, temporal dependencies, similarity among series or patterns and so on.

There are tools that perform the search for time series in a limited way, for instance, they just allow the user filter series selecting a category and a time period [Secretaria do Tesouro Nacional 2013]. Other perform search based on similarity among series using as input one time series - for instance, [Ding et al. 2008, Negi and Bansal 2005].

Still other approaches to search for time series are proposed by [Aßfalg et al. 2006, von Landesberger et al. 2009]. The former proposes a tool that searches time series using a range of values defined by the user as search parameter. The latter uses text in natural language, on financial databases. The disadvantage of these approaches is that they do not consider extra information that may be indirectly associated with series. Our work provides it by using semantics.

4. Characterization of the Contribution

This paper describes a software framework that helps users to find time series of interest by performing search via semantic annotations. The work of [Silva 2013] is used as the starting point to achieve this goal.

The proposed approach allows to solve the kinds of problems shown in Section 1. In order to return relevant results, besides text match, the search allows using extra information that may indirectly be associated to series. Figure 1 illustrates an overview

of the solution considered. In this framework, experts may add extra information to series semantically annotating parts of one or more time series. Series are stored in relational databases and may have many semantic annotations (step 1). Semantic annotations are stored in a RDF (Resource Description Framework) database (step 2) and are associated to external ontologies (step 3). It enables the use of Linked Data concepts to aggregate knowledge to annotations. Annotations are versioned through time.

Returning to the corn example presented in Section 1 and as shown in the figure 1, given a set of time series the researcher may use the string "milho" in the search for time series of interest (step 4). Navigating through ontologies that are connected to annotations, new information associated to "milho" is inferred, for example, "Zea Mays" (scientific name of corn), terms like "dentado" (that refers to texture of the corn grain), "Cercospora zeae-maydis" (fungus that affects corn), "mancha-branca" (corn disease), "espiga" (term related to corn), time period in which corn is harvested and places where corn is planted. Defining ontologies is very hard, so we intend to use known ontologies, like AGROVOC [1], which is an ontology that deals with several areas related with food and agriculture. As a consequence of using an ontology, the researcher will obtain as results (step 5) time series annotated with terms associated with "milho" (like "dentado"), temperature and humidity series for regions with corn, grain amount harvested on regions where "milho" is planted and dollar price series in the period in which "milho" is harvested. This will allow experts perform new types of correlations.

The framework supports collaboration among experts by allowing comments on the annotations (or meta-annotations). These meta-annotations are also structured; their values are a free-text values and they are associated to the versioned time series annotations. This new level of annotations records discussions over time, avoiding the repetition of time series annotations created just to register this collaboration (as made by [Silva 2013]). Therefore, new annotations will be created just when there is an agreement about the real change of content.

This framework is being validated with time series provided by EMBRAPA. These series contain Normalized Difference Vegetation Index (NDVI) values extracted from satellite images taken from Mato Grosso. Moreover, EMBRAPA series are textually annotated with information on the corresponding crop. We intend to expand our database with other kinds of time series – e.g., meteorological data. This will potentially allow retrieving heterogeneous series, given a set of terms, which are nevertheless indirectly (semantically) related. To evaluate this work, we will define a set of queries to be performed, to solve the problem described in Section 1, using semantic annotations in the search. Together with experts from EMBRAPA, we will compare this result to that provided by standard keyword based queries, to validate our proposal.

5. Ongoing Work

In order to achieve the main goal of this work, we first define a time series semantic annotation model (which we call TSSAM). Our model has the following components: (1) Subject: an URI (Uniform Resource Identifier) that represents one time series. This URI will be created by our software; (2) Annotation identifier: an URI that identifies one semantic annotation; (3) Annotation range: defines the granularity of the semantic

[1] http://aims.fao.org/standards/agrovoc/about

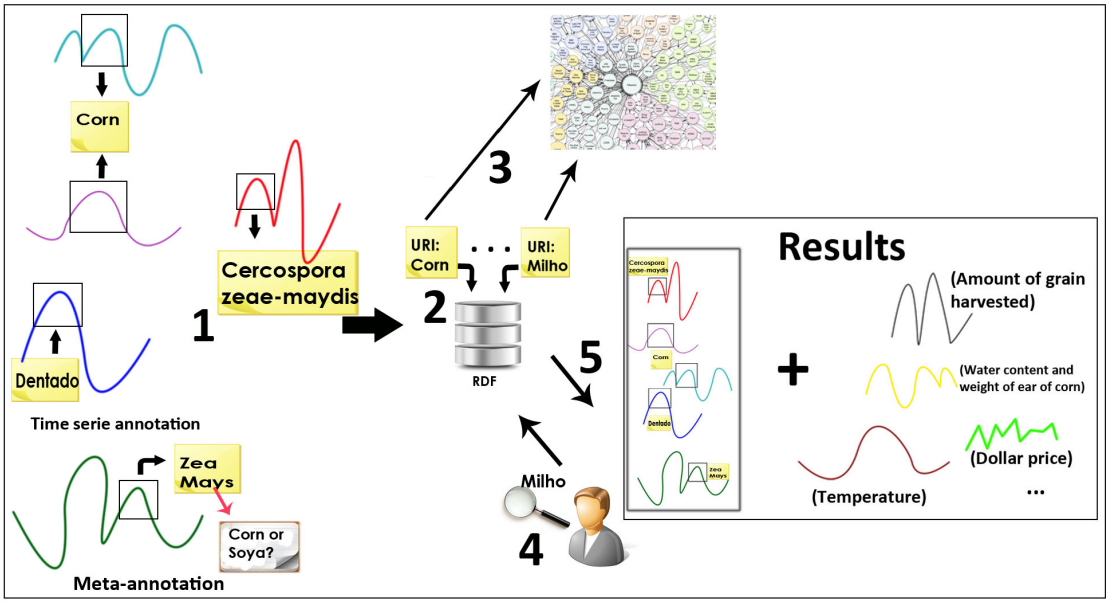

Figure 1. Overview of the model proposed

annotation, namely, a semantic annotation may be associated with all or part of one time series. This granularity will be specified by an time interval defined by the Time Ontology [2]; (4) Body: an URI representing the content of a semantic annotation. This URI is informed by the user; (5) Author: represents who creates the annotation; (6) Creation date: the date in which a semantic annotation was created; (7) Modification date: the date in which a semantic annotation was modified; (8) Version: version of a semantic annotation; (9) Time series property: represents a list of properties related with the annotated part of a time series, for instance, max point, amplitude, and so on.

Figure 2 illustrates our model in a RDF language where the ovals represent URIs and rectangles represent literals. The center of the figure contains the annotation identifier and around it are the other components of TSSAM. The "Time Series Property" is a rdf:Bag, which represents a group of literals. This model uses some properties already defined, for example, the property "annotates" relates an annotation to its resource and is defined in http://www.w3.org/2000/10/annotation-ns#annotates.

Once the TSSAM was defined, we developed an initial prototype. This first version enables the creation of one semantic annotation for parts of one time series, and one time series may have more than one annotation. As shown in figure 3, the user fills the form informing the URI that represents the annotated content (4^{th} TSSAM component), the annotation range (3^{rd} TSSAM component), and the author (5^{th} TSSAM component) in order to create a semantic annotation. The other components are automatically created by the software. On the right side it is possible to observe a graphical representation of semantic annotations for that time series and below more details about the annotations.

[2]http://www.w3.org/TR/owl-time/

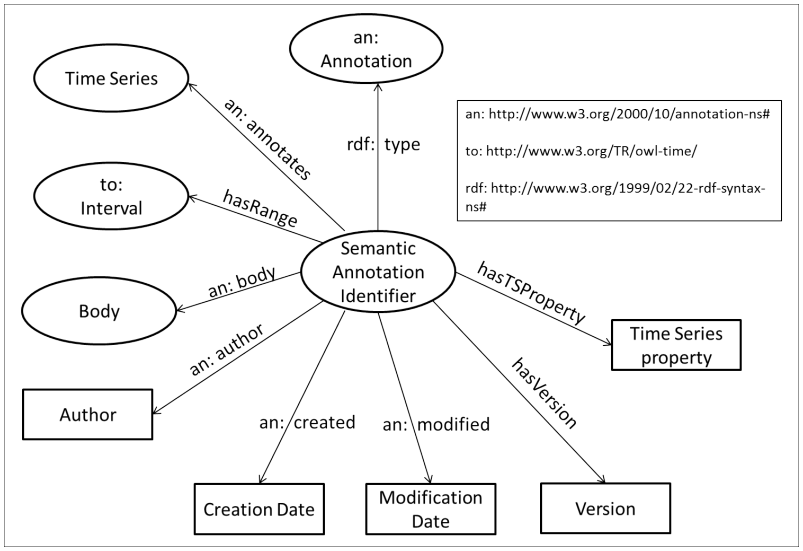

Figure 2. Time series semantic annotation model proposed

6. Conclusions and Future Work

This paper shows an ongoing research that helps experts share information, collaborate in production of data of common interest and search time series returning more relevant results. The framework proposed uses semantic annotations as a new possibility to search for time series, taking advantage of extra information attached to the series. As future work, we intend to improve our first prototype implementation and evaluate our semantic annotation model with experts from EMBRAPA. After that, the next stage is to develop algorithms to search time series via annotations, implement and evaluate them.

Acknowledgments Work partially financed by FAPESP (2014/07303-1), FAPESP/Cepid in Computational Engineering and Sciences (2013/08293-7), the MSR FAPESP Virtual Institute (NavScales project), CNPq (MuZOO Project), FAPESP-PRONEX (eScience project), INCT in Web Science, and grants from CNPq. We thank EMBRAPA for the data.

References

Aßfalg, J., Kriegel, H.-P., Kröger, P., Kunath, P., Pryakhin, A., and Renz, M. (2006). Tquest: threshold query execution for large sets of time series. In *Advances in Database Technology-EDBT 2006*, pages 1147–1150. Springer.

Ding, H., Trajcevski, G., Scheuermann, P., Wang, X., and Keogh, E. (2008). Querying and mining of time series data: Experimental comparison of representations and distance measures. *Proc. VLDB Endow.*, 1(2):1542–1552.

Gao, L. and Wang, X. S. (2009). Time series query. In Liu, L. and Özsu, M. T., editors, *Encyclopedia of Database Systems*, pages 3114–3119. Springer US.

Negi, T. and Bansal, V. (2005). Time series: Similarity search and its applications. In *Proceedings of the International Conference on Systemics, Cybernetics and Informatics: ICSCI-04, Hyderabad, India*, pages 528–533.

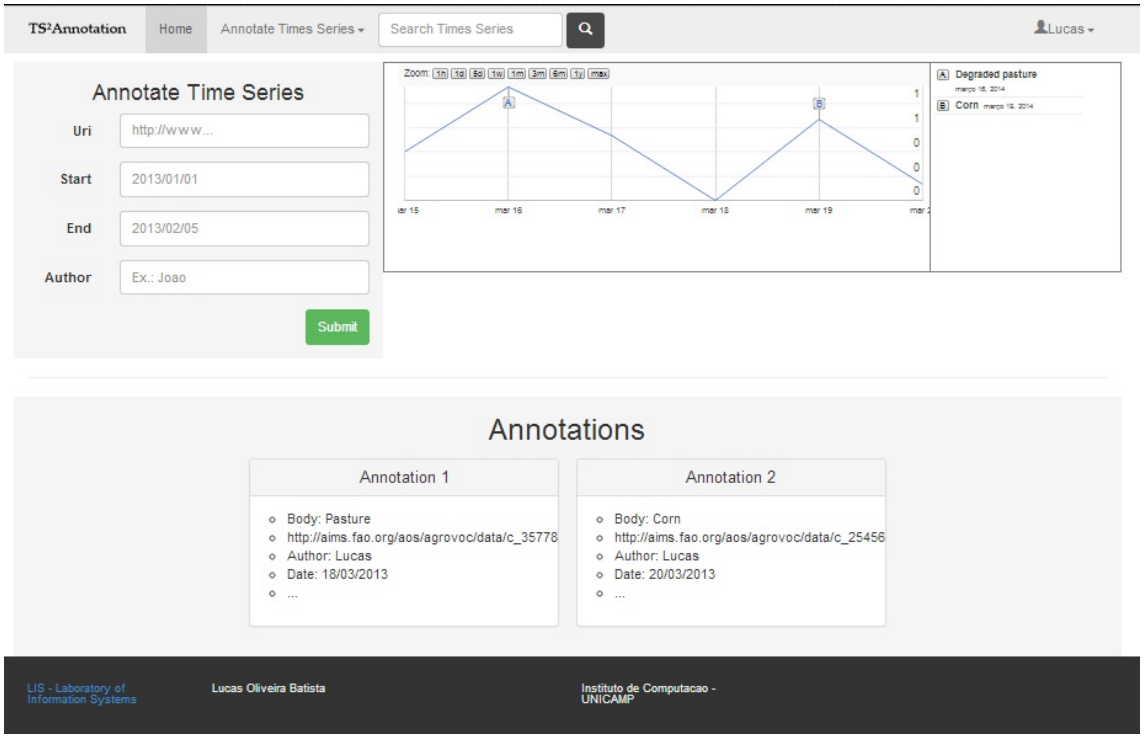

Figure 3. Time series semantic annotation prototype screen

Oren, E., Möller, K., Scerri, S., Handschuh, S., and Sintek, M. (2006). What are semantic annotations. *Relatório técnico. DERI Galway*.

Pariente, T., Fuentes, J. M., Sanguino, M. A., Yurtsever, S., Avellino, G., Rizzoli, A. E., and Nešić, S. (2011). A model for semantic annotation of environmental resources: the tatoo semantic framework. In *Environmental Software Systems. Frameworks of eEnvironment*, pages 419–427. Springer.

Pressly, Jr., W. B. S. (2008). Tspad: a tablet-pc based application for annotation and collaboration on time series data. In *Proc. of the 46th Annual Southeast Regional Conference*, ACM-SE 46, pages 166–171, New York, NY, USA. ACM.

Secretaria do Tesouro Nacional (2013). Séries temporais. `http://www3.tesouro.fazenda.gov.br/series_temporais/principal.aspx`. Acessado: 03-09-2013.

Silva, F. H. (2013). Serial annotator: Managing annotations of time series. Master's thesis, UNICAMP. Supervisor Medeiros.

Sousa, S. R. (2010). Gerenciamento de anotações semânticas de dados na web para aplicações agrícolas. Master's thesis, UNICAMP. Supervisor Medeiros.

von Landesberger, T., Voss, V., and Kohlhammer, J. (2009). Semantic search and visualization of time-series data. In *Networked Knowledge-Networked Media*, pages 205–216. Springer.