Supporting the study of correlations between time series via semantic annotations

Lucas Oliveira Batista[1], **Claudia Bauzer Medeiros**[1]

[1]Instituto de Computação – Universidade Estadual de Campinas (UNICAMP)
CEP 13083-852, Campinas - São Paulo - Brazil

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

Abstract. *This paper shows a work in progress to design and develop a software framework that supports experts in the correlation of time series. It will allow searching for time series via semantic annotations. Thereby fostering collaboration among experts, and aggregate knowledge to content.*

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 an annotation-specific languages and stored in files. In many domains, due to the fact that there are no standards, annotations have a large heterogeneity. This complicates the search, sharing and integration of data. 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 the label of a metadata field, and o is an ontology term that semantically describes this subject.

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. Given this scenario, this paper proposes a model of a framework to support the study of correlations between time series, where series are searched via semantic annotation. Besides allowing more sophisticated searches, and to aggregate knowledge to series, this framework will facilitate collaboration among experts.

2. Related Work

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

2.1. Time series annotation

Annotations may be made in many formats, for example, video, audio or text. This work focuses in textual annotations of time series. There are many tools that support series

annotation - for example, [Pressly 2008, Silva 2013]. Both enable to associate 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.

These tools present a few disadvantages. First of all, both 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.

2.2. 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 to select a series category and a time period [Secretaria do Tesouro Nacional 2013]. Others 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.

3. Model and methodology proposed

This paper describes a specification of a software framework that supports users to perform correlations among time series performing search via semantic annotation. The work of [Silva 2013] is used as the starting point to achieve this goal. This framework is validated with data provided by EMBRAPA.

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. There are challenges to find time series related to the problem, for instance, “corn” is cultivated in different parts of 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”, which complicates even more the obtention of results.

The proposed approach allows to solve these kinds of problem. In order to return other relevant results, the search for time series uses 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 kept apart from the time series themselves, aiming to use

less storage space as done by [Silva 2013] (step 2). Moreover, using Linked Data concepts, annotations are associated to external ontologies aggregating knowledge to content (step 3).

Returning to the example and as shown in the figure, the researcher may use a string "milho" in the search (step 4). Navigating through ontologies that are connected to annotations, new information associated to "milho" are inferred, for example, "Zea Mays" (scientific name of corn), terms like "dentado" (term 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], wich is an ontology covering several areas related with food and agriculture. It allows the subsequent search for time series with this new information. Thereby, the researcher will obtain as results (step 5) time series annotated with terms associated with "milho" (like "dentado"), temperature and humidity series, grain amount harvested on regions where "milho" is planted and dollar price series in the period in which "milho" is harvested. More accurate results and additional information about the content allow new types of correlations.

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

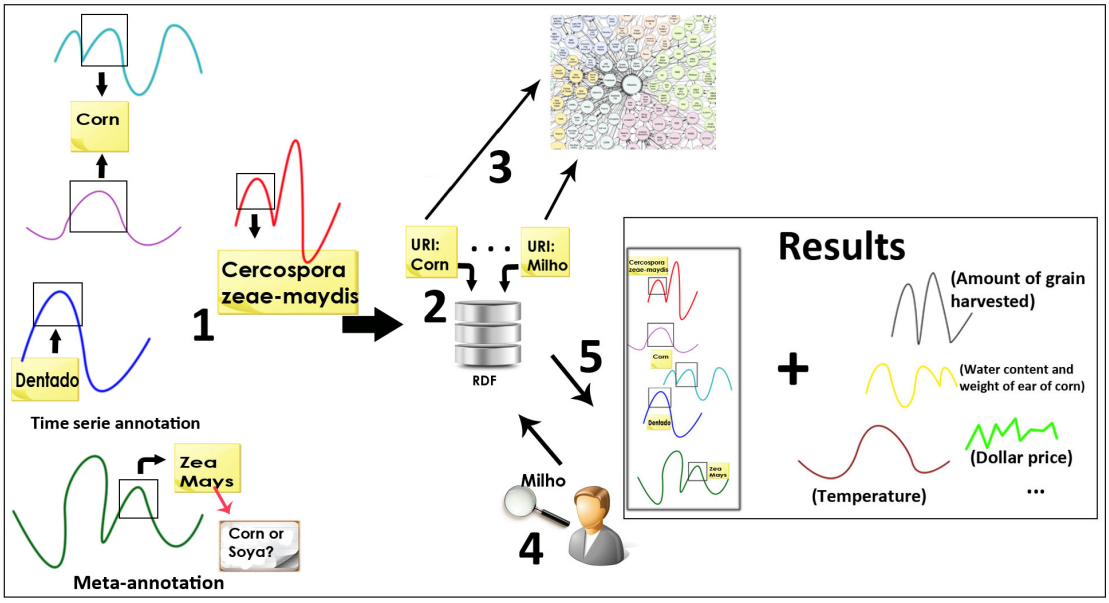

Figure 1. Overview of the model proposed

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

4. Conclusion and future work

This paper shows an ongoing research that helps experts to share information, collaborate in production of data of common interest and search time series returning more relevant results. The framework proposed uses semantic annotation as a new possibility to search time series, taking advantage of extra information attached to the series. Semantic annotations are interpreted by machines, allowing the use of automatic techniques on these data, for example, inference techniques. This new possibility expands and refines the search scope allowing more refined analyses to obtain results.

Acknowledgments Work partially financed by CAPES, FAPESP/Cepid in Computational Engineering and Sciences (2013/08293-7), the Microsoft Research FAPESP Virtual Institute (NavScales project), CNPq (MuZOO Project), FAPESP-PRONEX (eScience project), INCT in Web Science, and individual 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.

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

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, Universidade Estadual de Campinas - UNICAMP. Supervisor Claudia Bauzer Medeiros.

Sousa, S. R. (2010). Gerenciamento de anotações semânticas de dados na web para aplicações agrícolas. Master's thesis, Universidade Estadual de Campinas - UNICAMP. Supervisor Claudia Bauzer 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.