Keyword-Based Search over Environmental Datasets

José R.R. Viqueira[1], Alberto Bugarín[1], Joaquín Triñanes[2,3,4], Jaime Martínez-Urtaza[5].

1 Centro de Investigación en Tecnoloxías da Información, Univ. Santiago de Compostela.
2 Instituto de Investigacións Tecnolóxicas, Universidade de Santiago de Compostela.
3 Atlantic Oceanographic and Meteorological Laboratory, NOAA
4 Cooperative Institute for Marine and Atmospheric Studies, University of Miami.
5 Department of Biology and Biochemistry, University of Bath
{jrr.viqueira, alberto.bugarin.diz, joaquin.trinanes}@usc.es; j.l.martinez-urtaza@bath.ac.uk

Keywords: Keyword-based search, Environmental data management, Data mining, Semantic annotation, cholera, climate change.

1 Extended Abstract

The development of technologies that enable keyword-based search over structured datasets is a major challenge that is currently being faced by many researchers. Current research efforts in this area are mainly focused on entity/relationship (ER) business data, generally managed with relational technology and published with state of the art web and linked data standards (HTML, XML; RDF, etc.) [2,4].

Many open data efforts have been undertaken related to geographic and environmental domains, e.g. Spatial Data Infrastructure of the EU (INSPIRE) and the USA National Spatial Data Infrastructure. Many structured geographical and environmental datasets are available in the web, however, their specific characteristic make them unsuitable for the direct and effective application of the keyword-based techniques that are being developed for ER data. In particular, i) these datasets do not always fit ER data modelling, ii) they are generally large and highly heterogeneous and iii) general purpose environmental semantics may be considered, including data models [3], and ontologies [6].

Currently, geographic keyword-based search is supported over metadata by implementations of catalog services [5]. Some limited web search is also enabled for some ER environmental dataset[1]. Keyword-based search over environmental numeric datasets has not been studied yet, to the best of these authors knowledge.

The challenge proposed in the present document is the development of keyword-based search techniques for environmental datasets, which exploit both available metadata and also knowledge extracted from the numeric data. Thus, meteorological datasets could be queried using terms such as high winds, storm, hurricane, Katrina, tropical depression, hail, and blizzard. To achieve this, first, numerical datasets must be annotated with terms from well-known vocabularies such as SWEET [6]. The result of the annotation process will be a series of keyword indexes that will be used during the searching, ranking and browsing of the datasets.

---

[1] NOAA Storm Events Database: http://www.ncdc.noaa.gov/stormevents/

A number of Machine learning approaches can be applied in this context, ranging from conventional classification techniques to more advanced and linguistically interpretable approaches, some of them used in the "content determination" stage of Data to Text systems. Among the techniques capable of managing uncertainty in the labels, Computing with Words [7] provides a number of modelling tools (linguistic values and relationships, linguistic filtering of data) for detecting complex phenomena over numerical datasets or temporal data series. Its capability for representing fuzzy profiles of data (temporal evolution), spatio-temporal and other relationships among data, convert fuzzy-based approaches in a powerful tool for approaching the automatic annotation of databases when uncertain terms and vocabularies are involved.

Several science use cases can benefit from this approach. An example of this in the field of public health and water microbiology is related to the studies involving cholera infections. Cholera is a severe acute diarrheal disease with high mortality and morbidity rates in several areas of Asia and Africa, and devastating outbreaks in other regions (e.g. Haiti). Existing models rely on highly heterogeneous environmental, socio-demographic and geographic datasets, which include parameters such as rainfall, winds, sea and air temperatures, zooplankton, demography and economic development, among others. Many studies show that climate change will increase cholera incidence due to a number of factors, such as heat waves, ocean warming and heavy rainfall. The proposed approach would allow to quickly determining the datasets associated to a parameter, including the climatologies constrained by region and time interval. They can then be combined with epidemiological information to develop and implement new and improved risk models for the disease [1]. This use case will benefit of research being carried out by the authors focused on the investigation of the epidemic dynamics in Malaysia during recent years.

References

1. Baker-Austin C, Trinanes JA, Taylor NGH, Hartnell R, Siitonen A, Martinez-Urtaza J. 2012. Emerging Vibrio risk at high latitudes in response to ocean warming. Nat. Clim. Change 3:73–77.
2. S. Bergamaschi, F. Guerra, M. Interlandi, R. Trillo-Lado, and Y. Velegrakis. QUEST: a keyword search system for relational data based on semantic and machine learning techniques. Proc. VLDB Endow. 6, 12 (August 2013), 1222-1225. (2013).
3. S. Cox. Geographic Information – Observations and Measurements. OGC Abstract Specification Topic20. Open Geospatial Consortium (OGC). (2013). (http://www.opengeospatial.org/standards/om)
4. E. Demidova, X. Zhou, W. Nejdl. A Probabilistic Scheme for Keyword-Based Incremental Query Construction. IEEE Trans. Knowl. Data Eng., 24(3): 426-439 (2012).
5. D. Nebert, A. Whiteside, P. Vretanos, OpenGIS Catalogue Services Specification. Open Geospatial Consortium (OGC). (2007). (http://www.opengeospatial.org/standards/cat)
6. R.G. Raskin, M.J. Pan, Knowledge representation in the semantic web for earth and environmental terminology (SWEET). Comput. Geosci. 31, 1119–1125, (2005).
7. L.A. Zadeh. Fuzzy logic = computing with words. IEEE Transactions on Fuzzy Systems, 4(2):103–111 (1996).