Mining Knowledge TV: A Proposal for Data Integration in the Knowledge TV Environment

José Carlos Almeida Patrício Junior
Universidade Federal da Paraíba
João Pessoa – PB - Brasil
jcapjunior@gmail.com

Natasha Correia Queiroz Lino
Universidade Federal da Paraíba
João Pessoa – PB – Brasil
natasha@di.ufpb.br

ABSTRACT
This paper presents *Mining Knowledge TV*, a module for data mining that is part of the *Knowledge TV* (KTV) Project. KTV proposes the specification of a semantic layer that is embedded in a Digital TV (DTV) environment, improving the way that content is accessed by other applications.

Categories and Subject Descriptors
H.3.4 [**Systems and Software**]: Information networks

General Terms
Algorithms, Design, Languages, Standardization.

Keywords
Data Mining, Digital TV, Digital TV personalisation

1. INTRODUCTION
Interactive Digital TV [1,2,8] is a new stage of TV technology, which i*ntends to support the convergence of digital technologies* through a systematic change from analogical to digital equipments and infra-structure. This change generates modifications in the whole productive chain, mainly in the consumption of final content.

In this scenario, this paper aims at presenting the specification of Mining Knowledge TV- MKTV, which focuses on the integration of data mining [3] technology with semantic aspects, mostly of them derived from the AI Knowledge Representation and Semantic Web [4,5,6] research. The MKTV is being developed in the context the Brazilian System of Digital TV - SBTVD and is part of the project goal to give the TVDI a semantic layer. Among other aspects, it has the aim of providing a rich knowledge base of data descriptions, resources, services, applications and relations amount such elements.

2. MINING KNOWNLEDGE TV - MKTV
The main aim of MKTV is the implementation of a KDD environment, which focuses on data mining and semantic information on the Knowledge TV platform [7]. This solution will provide a priori unknown data to DTV applications that use the SBTVD Ginga middleware [8] , so that they can use this solution to face issues such as information overload, personalization, directed merchandizing and so on.

The mining process will be carried out on the data from many sources, mainly the sources that come from the *Service Information* (SI) metadata table, which uses the MPEG2 standard in the Ginga DTV environment. This standard is used to represent information about TV programs, services and multimedia interaction. Examples of such information are channels, program schedule, program classification, etc. User behaviour is also an important kind of information source because it indicates, for example, the channels usually watched with start time and total watching period. The useful content obtained by means of data mining will be semantically enriched through the use of ontologies and then provided as a service to NCL or Java languages application developers. This is possible because Ginga supports the development of applications using both languages on its architecture. More information about the Ginga architecture can be seen in [8].

The data mining process acts on all these sources and generates new information that is semantically enriched by means of a domain ontology. This semantic process enables a better analysis and turns more explicit the meaning of the data mining resultant discovered knowledge. This semantic is provided as a service and creates opportunities, which can be used for NCL or Java developers to implement more powerful and sophisticated applications.

3. ARCHITECTURE DESCRIPTION
3.1 Investigation of solution for data mining
Brazilian DTV is being characterized as an environment of technological convergence, new and extremely susceptible to changes. It is not yet completely standardized and it is constantly being updated. In this way, these aspects impose restrictions that we must consider during the architectural modelling. These evaluated aspects can be highlighted as restrictions:

- The small processing capacity of the set-top box;
- Reduced and unstable space for persistence of information;
- Mechanism for exclusion of applications when changing channels, i.e.; the change of channel will delete all application information related to that channel.

All these limitations in the architecture of the STB lead us to use a hybrid approach detached from the middleware. That means that the components of the KTV (and consequently the MKTV) with highest consumption of resources (such as processing power and memory) will be exploring the Ginga middleware return channel [8]. The return channel is the implementation of the htpp protocol on the DTV environment. That means that some components will be running on the web and will communicate via the Internet with DTV components.

3.2 Architecture
The *Mining Knowledge TV* (MKTV) is the component of the Knowledge TV architecture that accounts for the discovery and treatment of useful knowledge from the DTV data, users

behaviour and other sources such as the Web. These data are initially stored in a local relational database and gradually we will start the process of extraction, transformation and load (ETL) of information. After the ETL process, the data will follow for the next module that is the Data Warehouse (DW) [3], a technique that is commonly used in conjunction with Data Mining [3]. The DW will be organized in departmental Data Marts, in accordance with the domains and tasks to be mined (*e.g.* personalization, marketing, business), concentrating on historical data and integrated.

The historical data will be organised in the DW. Next, the Data Mining module applies data mining algorithms, searching and discovering useful patterns and information not known in the existent DW. The knowledge extracted through the MKTV will be encapsulated in semantic files with more expressive power (OWL files). Ontologies specification in OWL will be the standard for communication between the modules of the KTV. Figure 1 illustrates the KTV conceptual architecture and the MKTV module.

One application scenario is the problem of recommendation and personalization of content. To deal with such problems, specific modules, specified on the conceptual architecture, will be instantiated and executed. First the system stores the data that comes from the STB to a database. Then, the information related to user watched programs will be extracted to the Data Mart Personalization in the DW. After this process, it will be used clustering algorithms to find groups with similar preferences. Such knowledge discovered will feed and enrich the ontology specified in the semantic modelling layer and will return the pattern discovered in the form of recommendation to the user. For example, the next available programs similar to the ones the user uses to watch. Depending on the data mining goal, other tasks and algorithms can be applied to discover the desired knowledge.

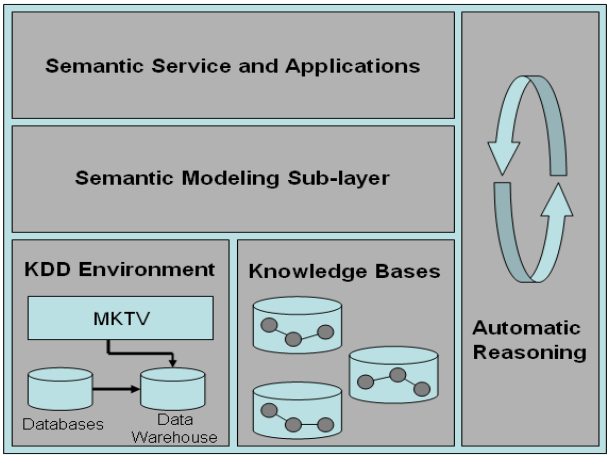

Figure 1. KTV Architecture and MKTV module

4. CONCLUSIONS AND DIRECTIONS

This paper describes our initial works on the *Mining Knowledge TV(MKTV)*, which is part of the *KTV* project. The major aim of this project is to provide semantic knowledge to be used for other DTV applications. At the moment, the MKTV is in a development stage, so that we have carried out a survey of the state of the art in data mining for DTV. In addition, we have also identified the main data mining methods and algorithms that are currently used in the DTV, together with a list of tools that are compatible to this new computational and interactive platform.

We can testify the innovation feature of this proposal if we consider the few DTV works that focus on the joint use of knowledge representation and data mining techniques to generate a better quality set of data.

The next MKTV activities intend to simulate DTV data traffic and integrate content from the data mining process and semantic modelling sub-layer. As future work, during its validation stage, MKTV will collaborate with the JCollab Project [16], whose aim is to develop a platform to create journalistic content via a social network. Another potential future work is the investigation about the integration of MKTV solution to other Digital TV systems.

5. REFERENCES

[1] Lekakos, G., Chorianopoulos, K., and Doukidis, G. 2007. Interactive Digital Television: Technologies and pplications. IGI Publishing. EUA.

[2] Lemos, G., Fernandes, J., and Elias, G. 2004. Introdução à Televisão Digital Interativa: Arquitetura, Protocolos, Padrões e Práticas. In: JAI Jornada de Atualização em informática. Salvador, Bahia, Brazil.

[3] Han, J., and Kamber, M. 2006. Data Mining Concepts and Techniques. 2a Edição, Editora Elsevier, UK

[4] Aroyo, L., Conconi, A., Dietze, S., Kaptein, A., Nixon, L., Nufer, C., Palmisano, D., Vignaroli, L., and Yankova, M. 2009. NoTube - Making TV a Medium for Personalized Interaction, EuroITV 2009, Leuven, Belgium.

[5] Yu, H., Dietze, S., and Benn, N. 2010. Semantic TV resources brokering towards future television. In 1st NoTube workshop on Future Television, in EuroITV 2010.

[6] World Wide Web Consortium. 2009. W3C Semantic Web Activity. (http://www.w3.org/2001/sw/)

[7] Lino, N., Araújo, J., Lemos, G., and Siebra, C. 2010. Aspectos Semânticos e Convergência Digital (Web e TV Digital). Proceedings of 2a. Conferência Web W3C Brasil (W3C Web.br 2010), Belo Horizonte, Brasil.

[8] Souza Filho, G. L. d.; Leite, L. E. C.; Batista, C. E. C. F.. Ginga-J: The Procedural Middleware for the Brazilian Digital TV System. In: Journal of the Brazilian Computer Society. No. 4, Vol. 13. p.47-56. ISSN: 0104-6500. Porto Alegre, RS, 2007

[9] Mangueira, J., Oliveira, F., Alves, K., Medeiros, A., Lemos, G. 2010. JCollab: Uma Ferramenta para Produção e Distribuição de Telejornais no Contexto da Web 2.0. In XXXVI Conferência Latino-Americana de Informatica – CLEI. Assunção – Paraguai