MIDAS SPORTS: A WEB DATABASE OF OLYMPICS LONG TERM RESULTS MANAGEMENT FOR BRAZILIAN ATHLETES PROSPECTION

Abstract ID: EASM-2015-196/R1 - (504)

All authors:
Fabio Matsunaga, Armando Toda, Jacques Brancher, Abdallah Achour Junior, Rosangela Busto

Date submitted: 2015-03-11

Date accepted: 2015-04-26

Type: Scientific

Keywords: athlete prospection
database
information extraction
physical documents
web crawler

Category: 14: Other sport management related issues

Synopsis:
Midas Sports is a web system composed of a database of long term sport events information storage, as athletes and results, and a web interface with summary projections about Brazilian athletes from different Olympic sport modalities and categories. All these information is textual data extracted from physical summary documents, using a module of document scanning and text mining, and from web, using a web crawler module to retrieve and extract automatically all textual data. All the sport information extracted is used to generate statistical summaries individual athlete performance analysis and their prospection.

Abstract:
AIM AND RESEARCH QUESTION
Sport results summaries are stored in physical documents kept in Olympic Committees. In long term, these documents are susceptible to physical damages over time, depending on storage and conservation conditions. This configuration leads to the needing of digital curation systems for documents preservation. Furthermore, even current and new data is stored on the Internet, web pages have lifespan, causing lost of older information.
Considering the exposed, the aim of this work is to propose a web database called Midas Sports, to store and group long term information related to Brazilian sport summaries from last 25 years, in order to aid long term performance of athletes and its prospection analysis, to detect possible future sport talents. All the sport information will be automatically extracted, integrating computational modules of information retrieval, data classification and extraction in textual data, considering two different resources: physical

files, which stores older data and web sites, which stores current data.

THEORETICAL BACKGROUND

In the last 45 years there is a great incentive in sports practice, specially for young people, making the manual management in spreadsheets very onerous (Arantes et al., 2012). Information Technology (IT) advances contributed to the digital information management in multidisciplinary applications, as the sport area, enhancing the service quality (Marefat & Faridfathi, 2015). In addition, web-based applications and knowledge-based systems contributed for people sport usage/habits sharing for social learning or competence improvement (Banias & Malita, 2011) and sport games results accurate forecasting or interesting results discovering using data mining system, due to the unreliability of simple objective and subjective methods for prediction (Leung & Joseph, 2014).

Sport related information have become accessible through digital media and resources, as and Official leagues championship or sports-based newspaper websites. These online resources contain summaries about events, athletes and results, available all the time and accessible from everyone, independent on the geographical location (Kuns et al., 2014). However, most of information are spread in different websites from different domains, which makes difficult to analyze the individual performance of an athlete, or even detect the possible or future sport talents. Considering the background, the focus of this paper lies on web system for athletes and results repository, aiming a management system development, mixing IT resources to contribute for sport science area.

METHODOLOGY

The physical papers data extraction consists on a document scanner module consisted on a FTP (File Transfer Protocol) server, from which the official summary documents were converted in digital image format and sent remotely to the Midas repository, from which textual data was extracted by optical character recognition (OCR) method. For the websites, a web crawler module runs constantly to search and filter the digital documents and web pages in official Brazilian Confederation pages, maintaining the database updated.

From the documents retrieved from both resources, all textual information were classified by Naive Bayes classifier, a supervised machine learning method, to categorize documents according to sports modalities and categories – races, hurdles, jumping sports, throwing sports and swimming. Natural Language Processing and rule-based methods were applied to obtain relevant information and organize it in parser trees, a computational data structure to efficiently manipulate and extract the respective data.

RESULTS, DISCUSSIONS AND CONCLUSIONS

The Midas Sports configuration enabled database information consultation by SQL (Structured Query Language) scripts for reports projections and statistical analysis. The projected database was linked in HTML (HyperText Markup Language) web pages using PHP (Hypertext Preprocessor) language and Google Charts API for web statistical graphics (histograms, pie charts and bar graphics) generation, which source code is embeddable in web pages.

The main contributions of this work is the availability of sport results from different modalities and categories in a single place, making a digital curation system. Some statistical analysis was possible to be performed, as the

performance of Brazilian runners from 100 meters athletics modality, the best performance of an athlete for each modality participated or how many times a particular athlete was among the first three places in the championships participated. All athlete performance can be filtered by State/gender/modality or even comparison between two athletes. Furthermore, it is possible to query a single athlete performance in different periods, since his/her early career to the present time, to determine if the athlete has a tendency to have good performances according to the extracted results, which will be evaluated by a professional sport area. This information is still being collected, but in the future, the performance data can be used to be compared with renowned and famous athletes to prospect and further identify new Olympic talents.

References:

Arantes, A., Martins, F. & Sarmento, P. (2012). Jogos escolares brasileiros: Reconstrução histórica. Motricidade, 8(2):916–924.

Banias, P. & Malita, L. (2011). Can we use sport, web 2.0 and social & informal learning to develop & enhance social competences? Procedia – Social and Behavioral Sciences, 15, 628–632.

Kunz, M. (2014). Micuim: Uma proposta de Sistema de Gerenciamento de Atividades Desportivas. EATI - Encontro Anual de Tecnologia da Informação e Semana Acadêmica de Tecnologia da Informação, 352–355.

Leung, C. K. & Joseph, K. W. (2014). Sports Data Mining: Prediction Results for the College Football Games. Procedia Computer Science, 35, 710–719.

Marefat, D. & Faridfathi, A. (2015). Relationship between Information Technology and Total Quality Management in Sport Federations. Journal of Applied Environmental and Biological, 5(3) ,52–58.