IJESRT

INTERNATIONAL JOURNAL OF ENGINEERING SCIENCES & RESEARCH TECHNOLOGY

MULTILEVEL DATA VERIFICATION ALERT SYSTEM FOR WRAPPER MAINTENANCE

Ms. Hema B. Waghmode*, Mr. A. A. Phatak, Mr. V. V. Pottigar
* Department of Computer Science & Engineering N.B.Navale Sinhgad College of Engineering, Solapur 413255

DOI: 10.5281/zenodo.557145

ABSTRACT

Unfortunately, websites are continuously evolving and structural changes happen with no forewarning, which usually results in wrappers working incorrectly. Thus, wrappers maintenance is necessary for detecting whether wrapper is extracting erroneous data. Wrappers are pieces of software used to extract data from websites and structure them for further application processing. The solution consists of using verification models to detect whether wrapper output is statistically similar to the output produced by the wrapper itself when it was successfully invoked in the past. Current proposals present some weaknesses, as the data used to build these models are supposed to be homogeneous, independent or representative enough, or following a single predefined mathematical model.

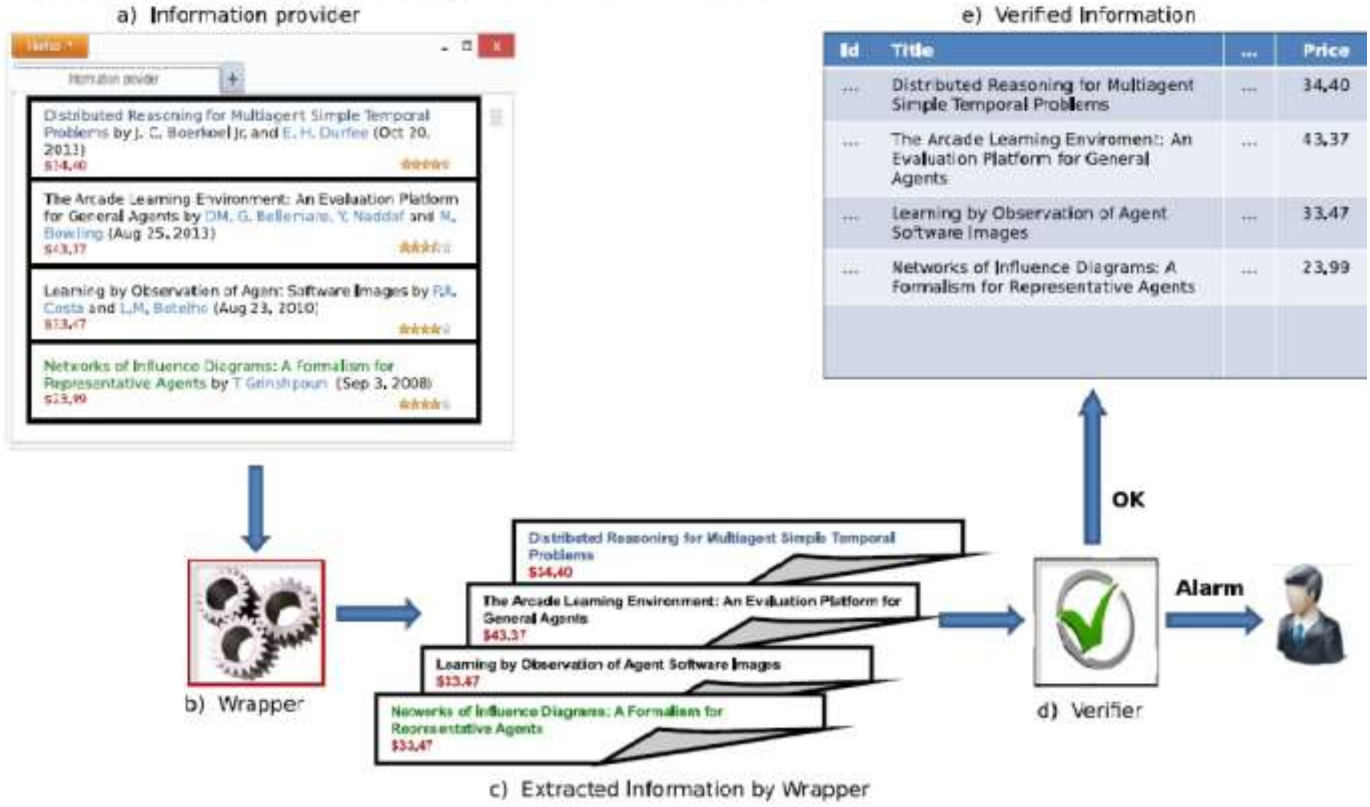
KEYWORDS: wrapper, classifier, multilevel.

INTRODUCTION

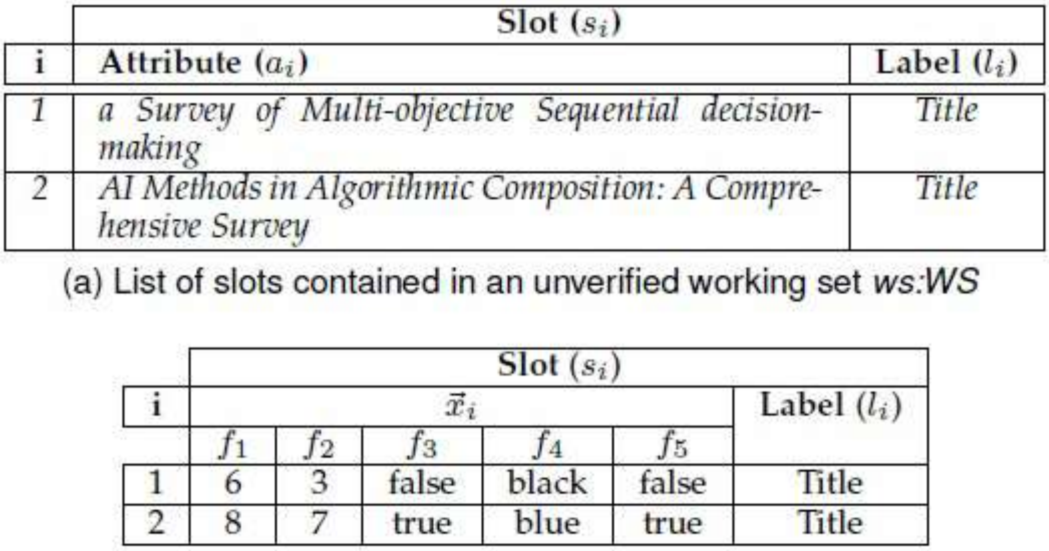
MAINTENANCE is a key challenge in the design of wrappers to extract and structure data from websites. Wrappers are used in a lot of real-world scenarios as enterprise information integration, context-aware advertising, database building, business intelligence and competitive intelligence, functional web application testing, opinion mining, or citation databases. Unfortunately, wrappers are not fully resilient to unexpected changes, as websites are dynamic entities that usually undergo changes in search forms, navigational models or the way information is rendered on the screen. The immediate consequence is that previously defined wrappers are no longer able to successfully extract data which results in system that manages corrupted or lost data. To illustrate the problem, Figure 1 presents a real scenario where a meta-search engine uses a wrapper (Figure 1.b) to extract and integrate the information residing in the website illustrated in Figure 1.a to provide business value added. The wrapper has been built using induction techniques to learn from a collection of manually annotated web pages as training data. Information extracted by the wrapper (Figure 1.c) is checked by a verifier (Figure 1.d). Then, imagine that the designers of the information provider decide to change the order of paper title and journal name presentation. This change would lead the meta-search engine to store inappropriate data or no longer able to store data at all.

ISSN: 2277-9655
Impact Factor: 4.116
CODEN: IJESS7

Fig. 1. Scholar metasearch engine service, integrating information from other site


a) Information provider
Distributed Reasoning for Multiagent Simple Temporal Problems by J. C. Boerkoel Jr. and E. H. Durfee (Oct 20, 2013)
$34,40
The Arcade Learning Environment: An Evaluation Platform for General Agents by DM. G. Bellemare, Y. Naddaf and M. Bowling (Aug 25, 2013)
$43,37
Learning by Observation of Agent Software Images by P.A. Costa and L.M. Botelho (Aug 23, 2010)
$33,47
Networks of Influence Diagrams: A Formalism for Representative Agents by T Geinshpoun (Sep 3, 2008)
$23,99
b) Wrapper
Distributed Reasoning for Multiagent Simple Temporal Problems
$34,40
The Arcade Learning Environment: An Evaluation Platform for General Agents
$43.37
Learning by Observation of Agent Software Images
$33.47
Networks of Influence Diagrams: A Formalism for Representative Agents
$33,47
c) Extracted Information by Wrapper
d) Verifier
OK
Alarm
e) Verified Information

Id	Title	...	Price
...	Distributed Reasoning for Multiagent Simple Temporal Problems	...	34,40
...	The Arcade Learning Environment: An Evaluation Platform for General Agents	...	43,37
...	Learning by Observation of Agent Software Images	...	33,47
...	Networks of Influence Diagrams: A Formalism for Representative Agents	...	23,99



i	Slot (s_i) Attribute (a_i)	Label (l_i)
1	*a Survey of Multi-objective Sequential decision-making*	*Title*
2	*AI Methods in Algorithmic Composition: A Comprehensive Survey*	*Title*

(a) List of slots contained in an unverified working set *ws:WS*

i	Slot (s_i) $\vec{x}_i$					Label (l_i)
	f_1	f_2	f_3	f_4	f_5	
1	6	3	false	black	false	Title
2	8	7	true	blue	true	Title

(b) Unverified training set

TABLE 2
Unverified data set

MATERIALS AND METHODS

Problem Statement

Extracting information from semi-structured Web pages is an increasingly important capability for Web-based software applications that perform information management functions, such as shopping agents and virtual travel assistants among others. These applications, often referred to as agents, rely on Web wrappers that extract information from semi-structured sources and convert it to a structured format. Comparison of relational form data in verifier is the challenging task.

OBJECTIVES AND SCOPES

Objectives

The idea of dealing with categorical and numerical features independently is to improve the verification process. Objective of this dissertation are:

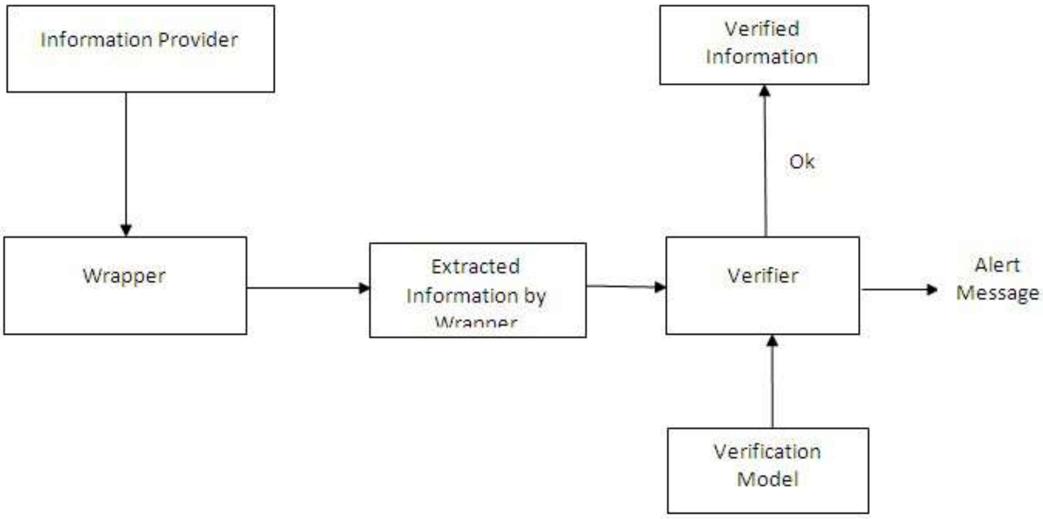
1. To deal with feature independence and identifies a subset of relevant features for use in model construction.
2. To deal with multiple verification models. Actually it generates as many models as roles, categorical and numerical features are used independently and, when profiling features, does not assume prefixed distribution.
3. Determines if a working set is invalid rely on the propriety of its records.
4. To prove that the proposed approach gives better performance than previous system.

Scope

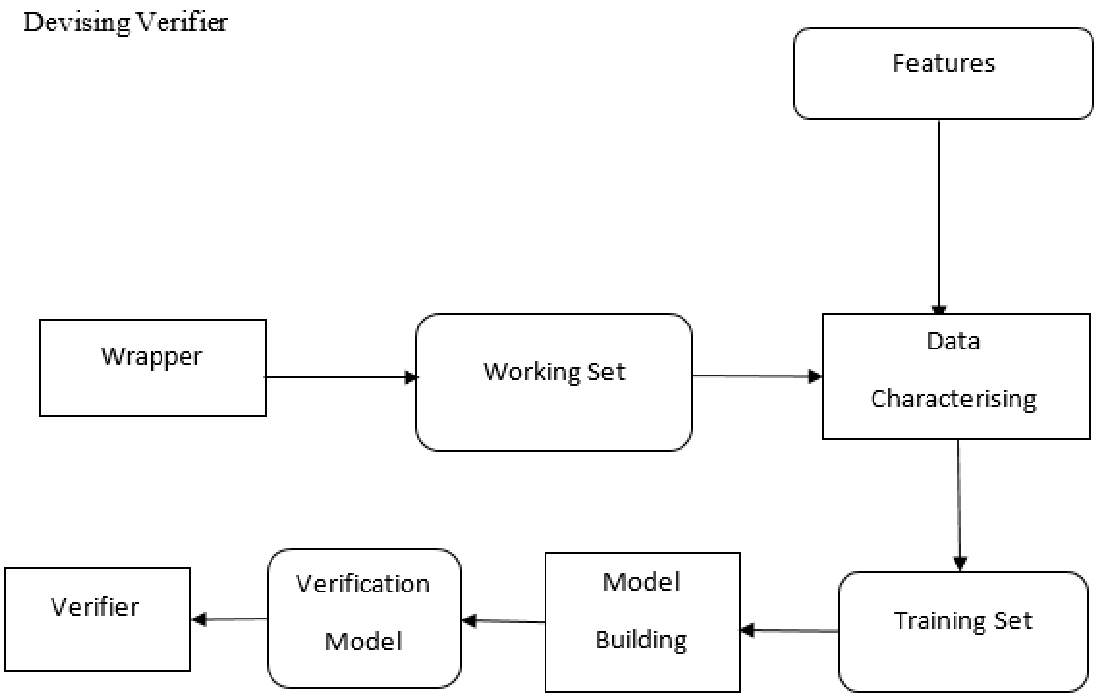
Proposed approach is used resolve the problems in existing wrapper verification system improve the overall performance of wrapper verification system.

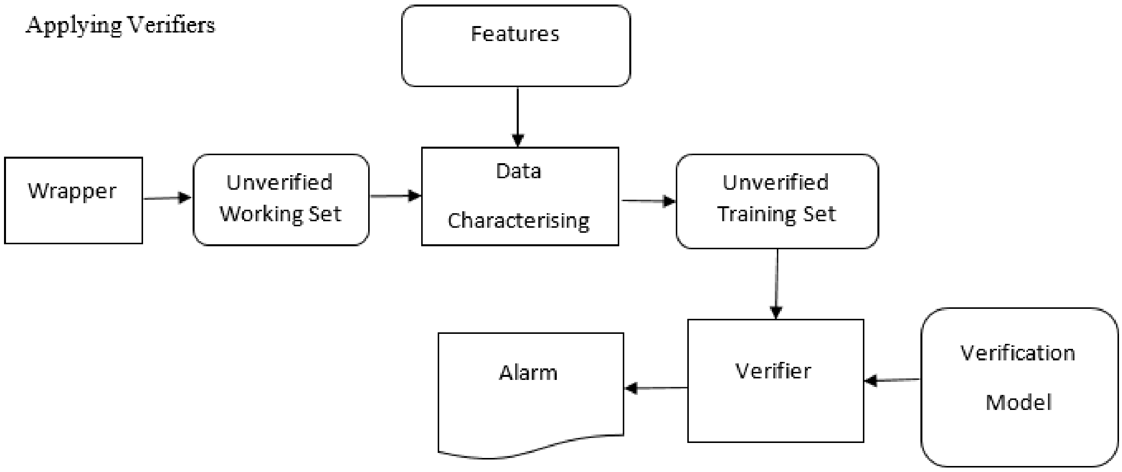
PROPOSED METHODOLOGY

MAVE (Multilevel wrapper Verification system), a novel multilevel solution to verify wrapper-extracted information. MAVE is based on two levels. At the first one, categorical features are used to generate a pattern, called signature which aim is to dismiss all elements that are regarded as non-valid. The second level only acts when the first one considers the element as valid, and it is the responsible of ratifying the validity by using standard One Class Classification (OCC) techniques.



This second level uses numerical features. OCC methodologists have emerged as techniques to solve classification problems in which one of the classes is well-sampled, whereas the others have very few instances or are not statistically representative. We prove that MAVE is well-suited for wrapper verification, overcoming weaknesses and achieving better results than current proposals. Its performance is evaluated by using the database. Then, non-parametric statistical analyse will be applied to compare the performance of MAVE and techniques enforced up to date. To solve the previous drawbacks, we present a new verification system based on One Class Classifiers. Contrarily to the previous proposals, ours is able to deal with heterogeneous working set but it is not able to generate invalid data sets.





CONCLUSION

A novel multilevel wrapper verification system to verify wrapper-extracted information is presented. This approach, named as MAVE, makes use of categorical and numerical features in two different levels of verification. Then, the idea of dealing with categorical and numerical features independently is proven to improve the verification process. Finally, MAVEs good performance relative to classical techniques acknowledged in literature is proven. Specifically, MAVE outperforms every technique used so far. The proposed approach take advantage of the idea that not only alert that wrapper is failing, but report the causes of failure in order to assist the wrapper maintenance. This would be possible because MAVE is able to identify the slot that is incorrectly labelled. Thus, we will try to identify the problem by analysing both the signature and the classifier output.

REFERENCES

[1] E. Ferrara, P. D. Meo, G. Fiumara, and R. Baumgartner,Web data extraction, applications and techniques: A survey, Knowledge Based Systems, vol. 70, p. 301323, 2014.

[2] P. Gulhane, R. Rastogi, S. H. Sengamedu, and A. Tengli,Exploiting content redundancy for web information extraction, Very Large Database Endowment, no. 1, pp. 578587,2010.

[3] T. G. Dietterich, Ensemble methods in machine learning,in International Workshop on Multiple Classi_er Systems,2000, pp. 115.

[4] V. J. Hodge and J. Austin, A survey of outlier detectionmethodologies, Arti_cial Intelligence Review, vol. 22, p.2004, 2004.

[5] V. Chandola, A. Banerjee, and V. Kumar, Anomaly detection: A survey, ACM Computing Surveys, vol. 41, no. 3,2009.

[6] M. Markou and S. Singh, Novelty detection: a review part 1: statistical approaches, Signal Processing, vol. 83, pp. 24812497,2003.

[7] C. D. He X and N. P, Laplacian score for feature selection,in NIPS: advances in neural information processing systems, 2005, pp. 810.

[8] T. C. Landgrebe, D. M. Tax, P. Paclik, and R. P. Duin,The interaction between classi_cation and reject performance for distancebased reject-option classi_ers, Pattern Recognition Letters, vol. 27, no. 8, pp. 908917, 2006.[9] J. Demsar, Statistical comparisons of classi_ers over multiple data sets, Journal of Machine Learning Research, vol.7, pp. 130, 2006.