International Journal of Advanced Scientific Technologies, Engineering and Management Sciences (IJASTEMS-ISSN: 2454-356X) *Volume.3,Special Issue.1,March.2017*

Analysis of Event logs in Internet by Using Process Mining Techniques

[1]S.Tejasree
Assistant Professor ,
Sri Venkateswara Engineering College for Women,
Karakambadi Road,Tirupati 517507
tejasree.s@svcolleges.edu.in

[2]Jasmine Sabeena
Assistant Professor ,
Sri Venkateswara Engineering College for Women,
Karakambadi Road,Tirupati 517507
jasminesabeena.s@svcolleges.edu.in

[3]M.Thulasi
Assistant Professor ,
Sri Venkateswara Engineering College for Women,
Karakambadi Road,Tirupati 517507
thulasi.m@svcolleges.edu.in

Abstract — This paper presents preparatory aftereffect of research venture, which is meant to consolidate metaphysics data recovery innovation and process mining apparatuses. The ontology depicting both information spaces and information sources are utilized to pursuit news in the Internet and to concentrate certainties. Handle Mining devices permits discovering regularities, relations between single occasions or occasion sorts to build formal models of procedures which can be utilized for the following resulting investigation by specialists. A relevance of the approach is considered with case of the ecological technogenic debacles brought about with oil slicks, and took after occasions. Ontologies permit conformity to new areas.

Index Terms — actuality extraction, occasion investigation, content mining, prepare mining, structure-focused data recovery.

I. INTRODUCTION

These days data recovery devices permit finding the productions containing data on occasions (actualities), on the items and relations, on the time when occasions happened and places thus on [9, 10, 11]. The approach permitting to uncover conditions between occasions (actualities), data on which is distributed in the Internet is offered in this paper.

The ontologies portraying both information spaces and information sources are utilized to pursuit news in the Internet and to concentrate certainties. The space ontology permit to set up association amongst articles and occasions, to arrange actualities, to execute a bunching, and so forth on the base of data distributed in the Internet. The ontology of information sources depict the data sources in the Internet. The consequences of data recovery are organized and put away in information base.

Occasion logs are shaped on base of the put away information in configurations utilized as a part of Process Mining devices. The information is cleared and nitty gritty at preprocessing with ontologies. Prepare Mining devices permits to discover regularities, relations between single occasions or occasion sorts, to develop formal models of procedures which can be utilized for the following resulting investigation by specialists.

At the examination model improvement existing devices of data recovery are utilized for occasions finding and logs creating and ProM framework is utilized for process mining.

Handle mining step by step infiltrates into the developing number of uses arrangement. The open door, which gives this teach, to find, screen and progress forms, beside such apparent reason as getting information from programming frameworks, are utilized for various errands like following and breaking down understudies' learning propensities in light of MOOC information, metaphysics driven information extraction from databases, in Workflow Management Systems for the medicinal services industry. This examination shows that consolidating Process Mining with different trains and methodologies can give an assortment of intriguing and nontrivial come about; this advances a ton of enthusiasm with respect to the logical environment.

International Journal of Advanced Scientific Technologies, Engineering and Management Sciences (IJASTEMS-ISSN: 2454-356X) *Volume.3,Special Issue.1,March.2017*

II. DEFINITIONS AND LIMITATIONS OF THE DOMAIN

As already mentioned, in our work we suggest to combine web, text and process mining in order to obtain un evident data patterns in a convenient and accessible graphical representation with the possibility of further model analysis. To analyze opportunities and demonstrate the described approach, technogenic accident subject area has been chosen, namely the events related to the oil spill. To understand what kinds of data can be collect from news feeds as part of this theme, a lot of query results were analyzed. Russian web-media and global search engines, like Google, were considered as data providers.

Further, assuming that the user may have two basic types of information needs:

1) Gross appearance in the industry, statistics;
2) Data on the specific event.

Two corresponding types of requests have been analyzed (f.e. the generalized – "oil spill", and a specific – "oil spill in Sakhalin April 5, 2016"). Obviously, the more general request is executed, the more diverse data we receive. So, by request for oil spill, results can be related to: the elimination of consequences, the sanctions measures, the new methodology to eliminate spills, actually oil disaster and many other types of events.

In this paper, we consider simple referential events:

- N. Samoilenko characterizes referential event as: "by event we understand the result of an action, behavior, occurrence, fact, which has a personal or social significance, something new, a change in the situation, the state of affairs".
- Simple events – internal form consists of the primary elements of event, first of all – action and associated

The minimal set of characteristics that we takes into account for events analysis are: participants of the event, geographical location, event border, internal relations between components of a single event, the relation between events. Such static event attributes as the company, the date format and geographical position will be a part of the domain ontology. Events borders were identified in the analysis of news feeds texts on the subject of oil spills. Since the news reports generally displayed the most important aspects of the case described, the event is considered as described in the ontology instance of the class "event" or, if there was no match in the ontology, any verbal constructions: verb + related words, met in the summary of the article. Communication components inside the event set during the semantic analysis that defines the verb structure to identify events. Dependency tree is constructed for each verb.

Further, there is the extraction of causal relations. This task is also facilitated by the analysis of news reports, since the size of the text is usually limited from two to five sentences, and the sequence of events often correspond to the stacking order. For escalating a chain of events, sequences extracted from various news reports will be linked by a given feature. News title does not belong to process able text, because it can significantly disturb the process of causal links identifying and duplicate information from the news.

However, the information from the header can be used to retrieve objects and event attributes if the event is duplicated in the news text. Also, an event in the header can be used as a marker for a situation by which a news-related situation will bind during further processing. In other words, by these markers, we can associate the events to traces. Trace in the Process Mining is a sequence of events, united by a common use case or a news message in our work. The event is an instance of activity that represents a well-defined process step. Traces are a display of the processes interaction.

III.FILLING THE DATABASE TO GENERATE MODELS

Clearly, on the particular demand we get the majority of a similar arrangement of occasions. On the off chance that this demand is created for a measurement model and changes over to a solitary follow for the general procedure, an issue emerges. The issue of data duplication in the follows can be settled by method for Process Mining. Nonetheless, synonymous in the information must be brought together with the utilization of cosmology.

Information recovered from Internet by means of ontology put away in the database in an indistinguishable path from client questions on which we were searching for news.

In the event that the client has created a demand, which things are not in our philosophy, we offer it to extend the metaphysics and may propose where he could include the idea. As a consequence of the work on this progression framework demonstrates the client settings from the philosophy that it can alternatively indicate (for instance, to confine the day and age, select an area, region, organization). At that point the framework produces the fundamental extra demands for news gateways, extricates writings, separates the data and fills the database.

All things considered, the client is again approached to choose what information parameters for the model he is keen on. For that the unique client ask for, removed information and the information put away in database are broke down. Condenses some normal virtual representation of the information identifying with every one of the parameters of the demand, on the premise of which the client is provoked to pick the criteria that he needs to be considered for the development of the procedure display. In view of this makes RDF petition for capacity and log era.

IV.USING PAGE STRUCTURE ONTOLOGY FOR INFORMATION EXTRACTION

In our approach the data look technique in view of web archives structure investigation and ontology usage is advertised. Two-level philosophy catching after portrayal should be created:

- Website (the web report being broke down source)

International Journal of Advanced Scientific Technologies, Engineering and Management Sciences (IJASTEMS-ISSN: 2454-356X) *Volume.3,Special Issue.1,March.2017*

structure depiction – primary page sorts and their interconnections;

• Web page data squares portrayal and their interconnections.

The illustration section of these two levels is delineated in the Fig. 1.

The primary level philosophy in essence keeps the portrayal of pages existing on the site being referred to – sitemap, however in more disentangled and bland shape. While building up a particular site portrayal, philosophy hubs will be populated with the locations of went to and broke down pages.

The second level philosophy is gone for keeping

the depiction of data pieces to be found on a site page, for example, route square, which can contain significant data, and additionally that of shaping these squares like frame controls, static or element pictures, tables, content regions etc. So as to build up this second level cosmology the most broad HTML layout sorts were analyzed alongside format arrangement components. While building up a particular website page portrayal, metaphysics hubs should catch markup puts unambiguously recognizing accurate positions of site page data hinders for further information extraction from it or from lower-level component constituting the square being referred to.

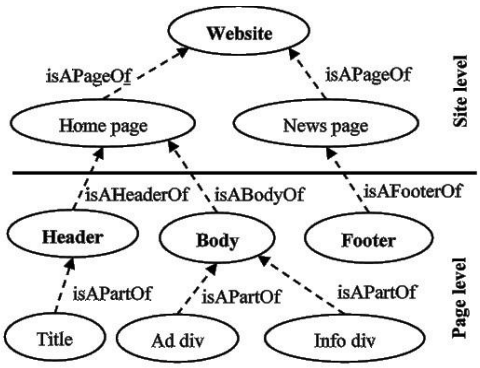

Fig. 1. Example Fragment of Two-level Web Document Ontology

In order to extract information from web documents by means of the proposed two-level structure ontology it is necessary to traverse it and identify the exact information block position for further content query.

Generally, there are several stages in the proposed ontology processing mechanism:

- Loading the ontology from local file or by the URI generated by the ontology editor exploited.
- Traversal algorithm execution and information block identification.
- Saving the placement address of the information division founded for further content query processing.

The main practical application to the offered method is to find more advantageous solution to the information extraction problem by boosting result relevance level paying more attention to the structure and placement of information.

The main advantages to this structure-centered information retrieval approach are as follows:

- web document can be annotated with the structure metadata allowing to take information placement into account;
- structure-centered information retrieval considers and exploits information divisions hierarchical structure interconnections;
- information placement metadata can help identify content duplication and filter it afterwards.

V. THE DATA STRUCTURE / DATA CLASSIFICATION / DATATYPING FOR STORING

In preparing the aftereffects of the above general kind of inquiries, notwithstanding the portrayed strides for every news thing or a follow, it is likewise important to order the circumstance depicted (as appeared in Figure 1, the outcomes for "oil slick" are distinctive sorts of occasions: social, political and natural). For every occasion class ought to be designated claim traits that can go about as markers relying upon the model that the client needs to get. Occasions are arranged as far as the characteristics that they may have. One news post may incorporate occasions identified with diverse classes. Key ascribes are applying to associate them. The accompanying fundamental sorts of occasions and key qualities were distinguished amid the examination of the news sustain of oil debacle:

✓ Disaster (date, oil organization, put) – straightforwardly catastrophe themselves, for example, fire, spill, blast.

✓ Financial suggestion (association) – evaluating the monetary harm, this incorporates as a cost for the disposal of outcomes too other financial pointers of endeavors, populace and nations.

✓ Industry news (oil organization, distribution date) – conceivable logical disclosures, accomplishments, developments in the field of oil industry, any data identified with the operation of organizations: growth, conclusion, insolvency.

✓ Sanction (date) – data on the authorizations and punishments.

✓ Socio-ecological ramifications (production date) – the effect on the populace, the casualties, harm to horticulture, the effect on society and the conceivable responses, showings, distress.

VI. MULTIPLE ADVANCED SEARCH QUERIES

✓ As already mentioned, once the data extracted by user request, they are analyzed and, depending on the completeness of the data set the user is able to:

- ✓ Expand the data set (Query for a specific event and query by attributes (geographical position, the company, the time interval)).
- ✓ Build a process models that are available at this stage.
- ✓ If the user decides to supplement the data with repeated request, the system provides the opportunity to extend the ontology, and using the concept of the updated domain ontology to generate a new request to the network news portals. Advanced data set allows creating more accurate and full covered process models.

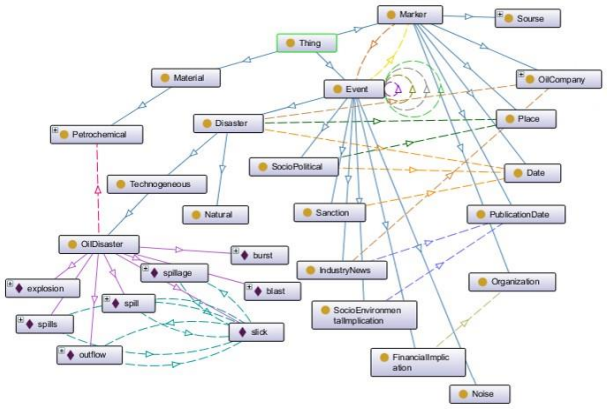

Fig.2. A fragment of the domain ontology example

Further, the concepts of the domain ontology (Fig. 2) are mapped to the ontology of the log (Fig. 3), and the output document in XES format creates. As a standard set of log extension used to describe data is not enough for our domain, it is necessary to expand. In addition to the below listed standard extensions, we need to include the date of publication, name of the organization that are not directly related to petroleum activities, the location of the event and the source of the news posts.

VII. USING THE CAPABILITIES OF STANDARD XES EXTENSIONS

In preparing the aftereffects of the above general kind of inquiries, notwithstanding the portrayed strides for every news thing or a follow, it is likewise important to order the circumstance depicted (as appeared in Figure 1, the outcomes for "oil slick" are distinctive sorts of occasions: social, political and natural). For every occasion class ought to be designated claim traits that can go about as markers relying upon the model that the client needs to get. Occasions are arranged as far as the characteristics that they may have.

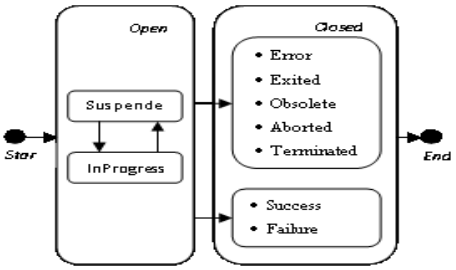

Fig. 3. Lifecycle transactional model

Further presented the model of the same process in the context of geographical position, i.e. request was extended by obtained in the previous step attribute Place = Mexico. In the model appeared some additional information on the socio-economic developments related to the assessment of damage and the imposition of a fine. Thus, clarifying and expanding the information requests the user fills the database with new data and gets more detailed models.

VII. REFERENCES

[1] M. Leemans; W.M.P. van der Aalst.ProcessMining in Software Systems: Discovering Real-Life Business Transactions and Process Models from Distributed Systems.

[2] P. Mukala, J. Buijs, M. Leemans, W. van der Aalst. Learning Analytics on Coursera Event Data: A Process Mining Approach. In: Proceedings 5th International Symposium on Data-driven Process Discovery and Analysis 2015. Pp.18-32.

[3] D. Calvanese, M. Montali, A. Syamsiyah, W.M.P. van der Aalst. Ontology-Driven Extraction of Event Logs from Relational Databases. In: Business Process Management Workshops 2015.

[4] R.S. Mans. Workflow Support for the Healthcare Domain. PhD Thesis. Technische Universiteit Eindhoven, Eindhoven, 2011.

[5] N.A. Samojlenko. Semantika sobytijnosti i sposoby ee vyrazheniya: avtoref. dis. kand. filol. nauk. Alma-Ata, 1991.

[6] V.E. Goldin. Imena rechevuh sobytiy, postupkov i zhanry russkoi rechi // Zhanry Rechi. – Saratov, 1977. – Pp.23-34.

[7] P.P. Maslov. obnaruzhenie i izvlechenie prichinno-sledstvennyh zakonomernostej iz teksta na estestvennom yazyke. In: Proceedings of Conference “Znaniya-Ontologii-Teorii”. 2009.

[8] Draft Standard for XES - eXtensible Event Stream - for achieving interoperability in event logs and event streams. The Institute of Electrical and

Electronics Engineers, 2016.

[9] V. Peña-Araya. Galean: Visualization of Geolocated News Events from Social Media / V. Peña-Araya, M. Quezada, B.

Poblete // Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '15). ACM New York. 2015. Pp. 1041-1042.

[10] M. Schuhmacher. Finding Relevant Relations in Relevant Documents. In: Advances in Information Retrieval: Proceedings of 38th European Conference on IR Research, ECIR. Padua, Italy, March 20-23, 2016. Pp. 654-660.