Semantic Formalization of Cross-site User Browsing Behavior

Julia Hoxha
Karlsruhe Institute of Technology
Institute of Applied Informatics and Formal Description Methods (AIFB)
Karlsruhe, Germany
julia.hoxha@kit.edu

Keywords
semantic log, cross-site browsing log formalization, classification of navigation logs, prediction with structured output, structured SVM

1. INTRODUCTION
Large amounts of data are being produced daily as detailed records of Web usage behavior, but the task of deriving knowledge from them still remains a challenge. Modeling and mining approaches are significant instruments to discover browsing patterns in such data and to understand how users browse Web sites.

There is an increasing body of literature on the investigation of clickstream data and navigation behavior modeling, with the majority focusing on data collected in a single site. Inspiring works [18] convincingly argue on the benefits of studying user behavior at multiple websites. Such approaches present significant potential to derive actionable behavioral knowledge and make better future forecasts, but they still have to tackle the problem of heterogeneity of the information encountered at different sites.

We approach the problem of usage data comprehensability at its root, addressing the issue of semantically formalizing cross-site user Web browsing behavior. Usage data (or usage logs) are syntactic representations of Uniform Resource Locator (URL) requests of pages and Web resources accessed by the site visitors. Due to the primarily syntactical nature of such requests, comprehension of users' browsing patterns is difficult. Hence, there is an urge for formalization approaches that leverage the semantics of the usage data in accordance with the domain they occurred.

As such, mapping usage logs to comprehensible events from the application domain helps to discover more insights about user behavior. While most approaches use flat taxonomies to represent such vocabulary, we deploy ontologies for structuring domain concepts and relations, since they ensure a richer semantic model of a Web site content.

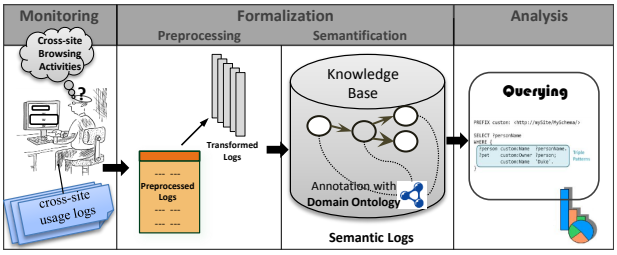

Figure 1: User Browsing Behavior Formalization Approach

This work aims at monitoring user behavior across multiple Web sites, logging clickthrough data upon agreement of Internet users. Each log entry is a tuple $\mathcal{L}$={*UserID, URL, timestamp*} of a user ID, URL of the accessed Web resource, and real time when this happened. These usage data logs are initially stored in raw form, as produced upon each user interaction.

The overall approach, illustrated in Fig. 1, comprises a series of steps, such as data preprocessing (human logs filtering, session construction, data transformation), formalization of usage logs, and techniques for their semantic enrichment.

This thesis will give the following contributions to the field:

- **Model for the formal and semantic representation of cross-site browsing logs**. I present the Web browsing Activity Model (WAM), expressed as an OWL-2-DL ontology, which enables a shared conceptualization of the knowledge from the various domains where the usage logs are recorded.
- **Techniques for the automatic extraction of the usage logs semantics** from heterogenous sources, in which the domain knowledge has a semi-structured or structured formal representation.
- **New approach for semi-supervised prediction with ontology-based output spaces**. This covers the problem of inferring the semantics of logs belonging to sites that do not offer a domain ontology. The contribution is a structured prediction algorithm formulated for the case of complex output objects rep-

resented as ontologies (with is-a hierarchies of classes and relations among them).

2. MOTIVATION, USE OR APPLICATIONS

Existing approaches can benefit from leveraging usage data with semantics in the following ways: increase understandability of user behavior with respect to the application domain; enable analysis on higher levels of abstraction e.g. for parameters in URL (museum instead of Louvre), or for location (capital instead of Paris), which can be also used for privacy protection; allow formulation of more expressive queries for mining user behavioral patterns. Furthermore, enrichment of usage data with domain knowledge provides a broader context of user behavior, which can be exploited for more intelligent recommendation models.

The semantically-leveraged logs provide an added-value with respect to their syntactic representation in being useful inputs for techniques such as semantic pattern mining, next-step navigation prediction or user clustering, which usually assume that the semantics of logs exists or are manually derived. A more beneficial aspect is the extension of these techniques to deal with cross-site browsing data and not only a single Web site.

2.1 Applications

An interesting application is the integration of domain knowledge in the process of discovering usage patterns. This helps to increase the precision, and hence interpretation of the retrieved patterns, while ensuring different level of abstraction. I present two approaches for discovering browsing behavior patterns, while using as basis the formal and semantic representation of logs:

I.Ontology-based Web usage mining
The first application deals with the automatic mining of frequent patterns from the sequence of event logs, which are enriched with description from domain ontologies. While recent trends in Web Usage Mining (WUM) have put the emphasis on the exploitation of ontologies to the pattern mining process, yet they share two limitations: the ontologies are either restricted to representations of class taxonomies while ignoring relates among the concepts, which reduces the problem back to the traditional generalized sequential pattern discovery [20], or they are restricted to a single ontology (single Site) that is assumed to be completed with relations. Because of the hetegoreneity of Web sites and respective domain knowledge, our setting requires a mining technique that addresses the problem when there are multiple ontologies in background and not all the relations among the concepts are established. Hence, they still need to be inferred during the mining process.

As a contribution to the WUM field, with practical motivation from the Web personalization field, I propose an approach for mining frequent sequential patterns in the presence of multiple domain ontologies. The mined patterns can, then, easily be extended to association rules [19], which provides predictions for the user's next step navigation preference.

II. Pattern discovery with $\mathcal{DL}$-LTL expressive queries
In this application, patterns are discovered from the corpus of the semantically formalized logs upon issuing specific queries that express semantic and temporal conditions of usage behavior.

While the first application (mining) concentrates on the semantics of the logs, another crucial aspect to consider when analyzing browsing behavior is also its temporal dynamic. Additional aspects of user browsing behavior can be discovered if reasoning not only with semantic constraints, but also with more expressive temporal conditions is made possible. I introduce an approach to formulate queries using a temporalized description logic called $\mathcal{DL}$-LTL, which combines $\mathcal{SROIQ}$ [8] with Lineal Temporal Logic (LTL) [1] over finite traces.

It is further shown how to search for behavioral patterns from the usage logs applying a query answering technique, which is based on current model checking tools. This allows to automatically retrieve sessions of user browsing events that satisfy a set of semantic and temporal conditions. The adaptation and application of the $\mathcal{DL}$-LTL logic and these techniques for the setting of Web usage analysis are novel.

3. STATE OF THE ART

The contributions related to this thesis are grouped into works dealing with 1) the modeling of user browsing behavior at multiple Web sites, 2) formal and semantic description of usage logs, and 3) exploitation of ontologies in Web usage mining, and 4) prediction of structured data.

3.1 Modeling Cross-site User Browsing Behavior

Interest to characterize online behavior has started much earlier with works such as those of Catledge *et al.* [5], and Montgomery *et al.* [16] that try to identify browsing strategies and patterns in the web. Browsing activity has been studied and modeled, e.g. Bucklin *et al.* [4] and others, usually exploiting server-side logs of visitors in a specific website.

Regarding the modeling of browsing behavior at multiple websites, Downey *et al.* [7] propose a state machine representation for describing search activities. The present an approach for modeling and analyzing user behavior, focusing on the search activities and what users do when they depart the search engine. Park and Fader [18] present a stochastic timing model of cross-site user visit behavior, using information from one site to explain the behavior at another. While, Johnson*et al.*[11] study online search and browsing behavior across competing e-commerce sites.

The works in this category do not particularly apply semantic techniques or ontologies for behavior modeling.

3.2 Ontologies in Usage Mining

There is an extensive body of work dealing with usage log analysis and mining, but we focus on the combination of these techniques with semantic technologies, which start with contributions such as Stumme *et al.* [22] and Oberle *et al.* [17]. In this field, research has been mostly focused on search query logs or user profiling. Recent approaches, which use semantics for extracting behavior patterns from

web navigation logs, are presented by Yilmaz *et al.* [27] and Mabroukeh *et al.* [14].

Vanzin *et al.* [26] present ontology-based filtering mechanisms for the retrieval of Web usage patterns. More recently, Mehdi *et al.* [15] tackle the problem of mining meaninigful usage patterns and exploit the impact of ontologies to solve this problem. These works are restricted to only one domain and not cross-site browsing behavior. Hence, they mostly deal with a mining problem in the presence of a single ontology. It is interesting to explore further the discovery of patterns when multiple domain ontologies are involved, considering the establishment of mappings between them as an additional requirement of the mining process.

It is important to note though, that the process of enriching of logs with semantics is not the central problem of these works. They mostly use the ontological knowledge in the background for leveraging or optimizing the mining techniques.

3.3 Semantic Formalization of Usage Logs

This group consists of works that directly deal with semantic annotation of usage logs, hence mapping the requests of Web resources to meaningful concepts from the application domain. **d'Aquin *et al.*[6]** present the UCIAD platform[1], which applies annotation of user-centric activity data. It relies on pre-defined URL patterns to characterize accessed resources over which the activities are realised, and therefore their respective semantics. As part of setting up the platform, it is initially defined which is the set of websites that are present on the considered server, as well as the URL patterns,expressed as regular expressions, enable to recognise webpages as parts of these websites. Similarly, definitions of the user activities are also manually made in the setup process, in order to characterize and give semantics to the user actions.

The work of **Tvarozek *et al.* [25]**, while actually focusing on an architecture for the personalized presentation layer of Web-based information systems, covers in one of its techniques the problem of semantically annotating usage logs. In order to create comprehensive logs of user actions, the logs browsing events captured by a client side monitoring tool, as well as server-side logging data, are enhanced with semantics from the Web sites content using a SemanticLog tool. This tool is based on a semantically-enabled portal, which means that there is a conceptual ontology in the background of the site. The mapping of an interaction of the user with parts of the Web site, then use this ontology to generate the annotation of the user action. In this case, the semantics of the logs are not inferred, but rather defined in background as part of the engineering of the site. Still, this can be feasible only in the case when one is in charge of the content of the site, and also restricted to a small set of sites. Additional manual effort in the engineering process is needed to generate the semantic annotations.

Stühmer *et al.*[21] focus on processing complex events of user interactions with annotated Web pages, and they also present an approach for capturing and lifting these events in RDF. Hence, instead of dealing with the syntactical form of events, they also address leveraging logs with semantic information, which pertains to the actual domain knowledge of the Web page. As in the previous work, this technique also assumes the presence of a semantically-enabled Web site. In this case, RDFa is used to support the semantics embedded within actual Web page data and allow reusable semantic markup inside of Web pages.

The related works in this group are restricted to a manual approach for enriching the logs with semantics. This limitation poses a significant burden when we need to analyse browsing behavior at various Web sites, which leads to immense efforts of extracting the semantics of logs and mapping them to respective domain ontologies. Moreover, it is assumed that the domain ontology is provided. This leaves the problem of inferring (learning) the semantic types of logs for non-semantically enabled sites still a challenge.

3.4 Prediction of Structured Data

Machine Learning today offers a broad range of methods for classification and regression, but only a few cover the problem of predicting complex objects, such as trees or graphs. The approaches dealing with prediction of structured and interdependent output data are principally grouped into those using probabilistic models (e.g. Conditional Graphical Models, HMM) and those using discriminative models (e.g. Max-Margin Structured Classification [23], Energy-Based Models [13], SVM).

In the latter group, Support Vector Machines (SVM) for structured and interdependent output spaces [24, 10] offer solid theoretical foundations, as well as very high efficiency for the structured prediction approach. While structural SVMs provide a generalized formulation of the learning problem, its state of the art applications cover only the case when the output object are sequences or trees.

There is still the need to reformulate the learning problem, and further adapt the SVMs for the case when the output instances are objects represented as ontologies. In this case, ontologies comprise not only a hierarchical structure of the classes (is-a hierarchy) in the output space, but also a set of semantic relations between these classes. The difficulty of the prediction problem now increases, since it requires learning a model that takes into account the semantics of the ontology in the output space, which is an additional requirement when compared to the current techniques that deal with general graphs or trees.

4. RESEARCH CONTRIBUTION

What is novel and innovative in the thesis?

I) a model to formally and semantically structure usage logs,
II) an approach for the automatic extraction of the semantics of usage logs from heterogenous sources, in which the domain knowledge has a semi-structured or structured formal representation, and
III) a new approach for semi-supervised prediction with ontology-based output spaces

[1] http://uciad.info/ub/

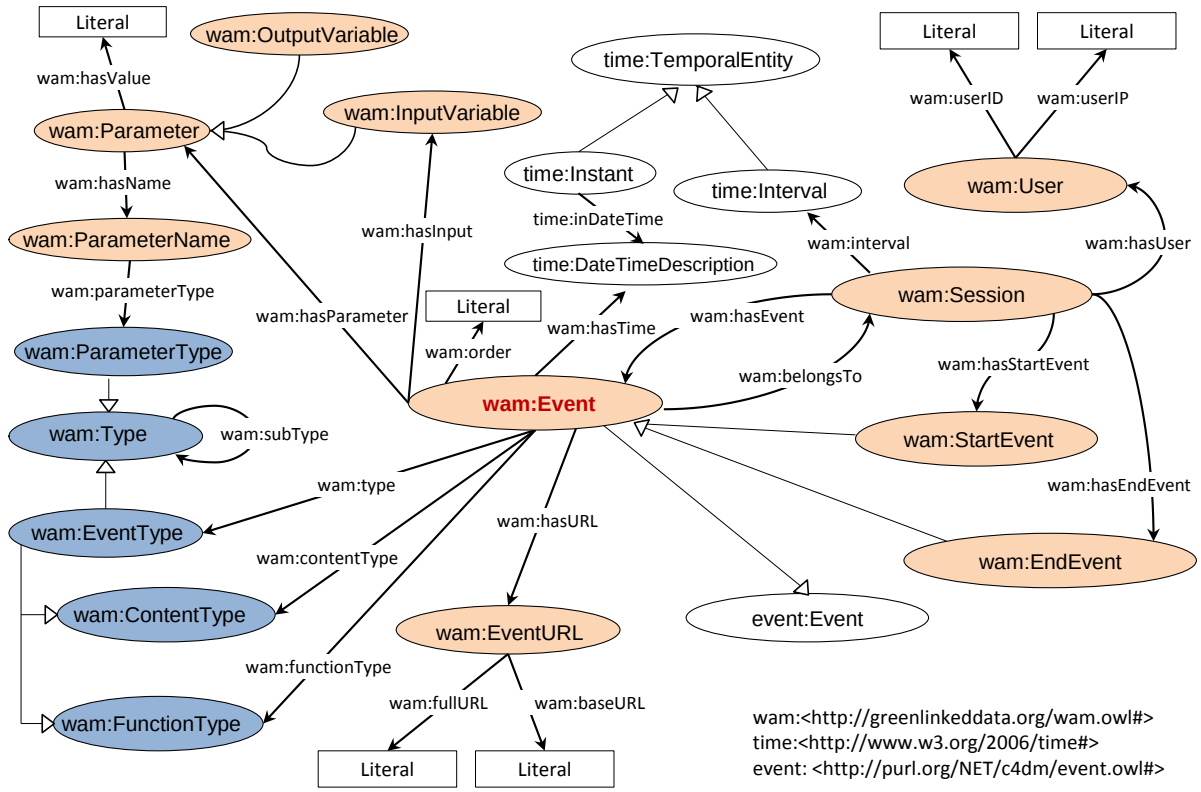

Figure 2: WAM Ontology

4.1 Formal Model for the Representation of Logs

I use the term *browsing event* to describe the basic component of user behavior in performing activities (actions) with the Web browser directly. Each event resulting from the interaction of a user with a specific Web page serves a particular function (searching within a portal, searching in a search engine, browse information, booking, login, etc.) related to some content (e.g. flight reservation, car rental, organization, person, hotel, et.c). Furthermore, events are grouped into sessions, which represent a period of sustained Web usage.

I use a set of definitions related to the concepts Event, Session, etc. For the realization of these concepts, I have use a Web Browsing Activity Model (WAM), which I formalize as an ontology (Fig. 2). This is also presented in the paper Hoxha et al. [9]

Classes and Properties. Classes in WAM are divided in three groups: Core classes, External classes, and Type classes. External classes are basic concepts that I reuse from well-established ontologies. Each `wam:Event` is a subclass of the concept `event:Event` from the *Event ontology* [2].

Each `wam:Session` has one `wam:StartEvent` and one `wam:EndEvent`, both of type `wam:Event`. Class `wam:User` is simply characterized by user IP address and ID, but the ontology allows flexible future extendability with user profiles or other attributes (e.g IP-based geographical location). To annotate of event timestamps and session interval, I reuse basic concepts from *OWL Time ontology*, [3] which models knowledge about time such as temporal units, instants, etc.

The ontology is expressed in OWL-2-DL with underlying $\mathcal{SROIQ}$ logic [8].

4.2 Automatically Extracting Log Semantics from Heterogeneous Sources

One of the contributions of this work is an approach for automatic extraction of logs semantics from the knowledge of the application domain where these logs were recorded. This approach covers the cases when the domain knowledge is structured as a formal ontology, and when it is represented in HTML-embedded RDF formats.

An example of enriching an event with semantics is illustrated in Figure 3. In its initial form, log consists of a syntactic representation of the URL request (in this case a demo paper). I retrieve the respective RDF representation and identify the accessed resource via its URI. Querying (via SPARQL[4]) the domain ontology, here the *SWRC* publications ontology, I can enrich the eventt's semantics with additional knowledge. I find that this resource is a *Demo* of type *InProceedings*. I can further extend the context with information like the *conference* WWW2011 it belongs, the

[2]http://purl.org/NET/c4dm/event.owl#

[3]http://www.w3.org/2006/time#

[4]http://www.w3.org/TR/rdf-sparql-query/

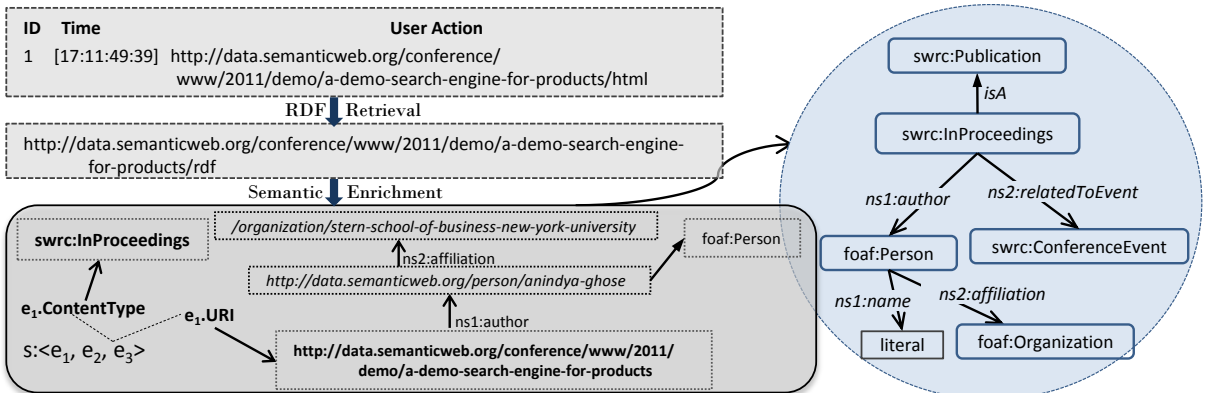

Figure 3: Semantic Enrichment of Usage Logs

conference *location*, the *author* of the paper, the authort's *affiliation*, etc.).

4.3 Semi-supervised Prediction with Ontology-based Output Spaces

The problem addressed in this work is the semi-supervised learning of the semantics of usage logs. The task is to predict, based on observations from training data, the classes to which the browsing events of the logs belong. Accordingly, I aim to infer the respective classes and relations between them, which comprise knowledge of the application domain structured as a formal ontology.

This work is based on the generalized formulation of learning with structural Support Vector Machines (SVM) [24], which involves features extracted jointly from input and output spaces. In this case, we are interested on the problem of learning a function $f : \mathcal{X} \rightarrow \mathcal{Y}$, which maps input instance $\mathbf{x} \in \mathcal{X}$ to discrete outputs $\mathbf{y} \in \mathcal{Y}$ based on a training sample of input-output pairs. In our setting, input space $\mathcal{X}$ consists of the browsing event sequences, whereas the output space comprises hierarchical structures of the `contentType` classes and semantic relations between these classes.

Structural SVM offer the capability of learning such a function for a structured, interdependent output space.The problem addressed is to learn a discriminant function $F : \mathcal{X} \times \mathcal{Y} \rightarrow \Re$ over input/output pairs, so that for a given input $\mathbf{x}$, we can derive a prediction by maximizing F over the response variables. Hence, one has to find f parametrized by a weight vector $\mathbf{w}$ such that:

$$F(\mathbf{x}; \mathbf{w}) = \underset{\mathbf{y} \in \mathcal{Y}}{\mathbf{argmax}}\ F(\mathbf{x}, \mathbf{y}; \mathbf{w}) \qquad (1)$$

The discriminant function F can be considered as a $\mathbf{w}$-parameterized family of cost functions, defined in a way that the minimum of $F(\mathbf{x}, \bullet; \mathbf{w})$ is at the desired output $\mathbf{y}$ for inputs $\mathbf{x}$ of interest. F is linear in a particular **combined feature representation** of inputs and outputs $\Psi(x; y)$,

$$F(\mathbf{x}, \mathbf{y}; \mathbf{w}) = \langle \mathbf{w}, \Psi(\mathbf{x}, \mathbf{y}) \rangle \qquad (2)$$

The specific form of Ψ depends on the nature of the problem, and needs to be accordingly designed for our task.

I intend to design a learning approach that uses these fundamentals, but whose formulation is extended to satisfy the following additional requirements:

- **output space is structured and interdependent**: the output is structured as an ontology (not just hierarchical representation), which consists of classes and relations modeling the Web sites domain knowledge. This requires the definition of: 1) the joint feature mapping Ψ, 2)appropriate loss function, and 3) algorithms for solving argmax problems of prediction- and loss-augmented inference.
- **presence of unlabeled data in the training dataset**: shifting the problem from a supervised to a semi-supervised approach. I explore two different ways to deal with the presence of unlabeled data: 1) develop a model find and propagate labels in training data before the learning step, and 2) follow a co-training approach.

Additional attributes characterizing the prediction problem is illustrated in Fig. 4, in terms of the types of input and output involved, features that are to be considered, and characteristics of labeled training data.

Part of the methodology in solving the learning problem is the theoretical argumentation for selecting SVMs as the basis of our approach, and accordingly, the exploration of comparable algorithms (e.g. Conditional Graphical Models, Max-Margin Markov Networks, Energy-based Models).

There are several reasons for choosing structural SVM as our classification approach. Firstly, SVMs in general are shown to perform better in building complex and accurate models [10], particularly in settings similar to ours such as Web page categorization or purely URL-based page classification [12, 2]. Secondly, SVMs deal very well with sparse and highly dimensional data, as is the case of the huge and heterogeneous amounts of cross-site usage logs, which lead to feature vectors that are large and highly sparse. At last, structural SVMs enable learning for complex and interdependent objects of the output space, leading us towards an extension of our approach in learning a formal, structured ontology with class relationships for the classification of events (i.e. requested resources) in the usage logs.

Input	Features	Output	Labels
Sessionized web logs: – Query log – Browsing log Logs from: – Various websites – One topical domain: 1) scientific publications 2) travel arrangement	1) URL Bag-of-Words representation: – Token – N-grams – All-grams 2) URL components: - Domain - Parameters 3) URL structure: - Sequential token dependency - Precedence Bigram 4) Sequential information – Path prefix – Query parameters in path	– Hierarchical structure + class relations – Formal Ontology	Labeled & unlabeled data in the training set Unbalanced classes

Figure 4: Characteristics of the prediction problem

4.4 Research Questions

1. How to preprocess the huge amount of logs gathered from users's browsing activity in multiple Web sites?

2. Which schema to use to describe logs of navigation at various domains in a formal and semantic way, hence, provide *semantic logs*?

3. How to map syntactic logs to meaningful events of the application domain where they were issued, i.e. how to enable log semantic enrichment in the case of various heterogeneous domains, and not a single site?

4. How to extract the semantics of logs in the case when the Web sites already provide some structured form of their domain knowledge (e.g. formal ontology, HTML-embedded RDF), taking into considerations that each domain has its independent knowledge model?

5. How to infer the semantics of logs belonging to the multiple sites that do not offer a domain ontology? How to classify these logs, therefore, learning their semantics, in a semi-supervised way?

6. How to evaluate the overall formalization approach: which new metrics to develop; which experiments to design and perform, preferably with real-word datasets?

7. How to demonstrate the benefits of the formalization approach based on applications/use case scenarios that profit from the semantically-leveraged logs?

5. EVALUATION

The aspects of the work that need to be evaluated are: 1) the completeness of formal semantics extracted from the background domain knowledge of the Web sites, 2) the accuracy of predictions done in the classification approach, in which the semantic types of the logs are learned in a semi-supervised way.

The evaluation of the approach dealing the automatic extraction of log semantics comprises a thorough quantitative analysis in terms of the degree of completion of the logs with semantic types (event content type, parameter types) retrieved from respective domain knowledge. As for the problem of predicting the semantic types, there is a necessity for an extensive accuracy-based evaluation, preferebly on real-world datasets of usage logs. In this part, new metrics need to be designed to evaluate the approach. The evaluation measure will be an harmonic mean of precision and recall, but adapted to the case when we have multiple classes in the output space, therefore an averaged measure is necessary, furthermore addressing the presence of parent-child and general relations among the classes.

A part of the evaluation will be conducted with datasets provided from USEWOD [3]. The USEWOD datasets consist of server logs from from four major web servers publishing datasets on the Web of linked data. In particular, I will consider the datasets containing logs from:
DBPedia: slices of log data spanning several months from the linked data twin of Wikipedia, one of the focal points of the Web of data. The logs were kindly made available to us for the challenge by OpenLink Software! Further details about this part of the dataset to follow.
SWDF: Semantic Web Dog Food is a constantly growing

dataset of publications, people and organisations in the Web and Semantic Web area, covering several of the major conferences and workshops, including WWW, ISWC and ESWC. The logs contain two years of requests to the server from about 12/2008 until 12/2010.

Nevertheless, the USEWOD datasets are very restricted with respect to the presence of logs at multiple Web sites. Hence, a crucial part of the evaluation will consists of experiment results gathered from datasets, which comprise real cross-site browsing logs. In this case, it is important to estimate the performance of the learning approach when experimenting with features that explore the sequential information of logs across heterogenous sites, which is also the kernel of the problem.

5.1 Application-based Evaluation

For the overall semantic formalization approach, another way of showing its added-value when compared to the syntactical representation of logs is the evaluation in two practical use cases. For this reason, I will present and implement two applications, in which the logs are leveraged with semantics: I) discovery of semantically enriched usage patterns and II) prediction of the usert's next navigation step.

In both applications, the baseline for the evaluation is the case when logs are not leveraged with semantics from the domain. For the first application, the main aspect of the evaluation is the increase in the expressiveness of usage patterns when the original logs of user behavior encompass later more semantic information that was extracted from the domain knowledge, or learned from the overall data. This scenario will involve a qualitative evaluation approach, in which the discovered usage patterns are manually evaluated in terms of their correctness.

For the second scenario, I will evaluate based on metrics such as precision and recall the impact that the semantically enhanced logs have on the general performance of predicting the next navigation step that the user might be interested to perform. Again, the baseline for the evaluation is the case of the syntactical representation of usage logs. This is a qualitative evaluation performed on labeled training datasets, which we plan to derive from real-world logs that will be partly manually annotated for the training part.

6. REFERENCES

[1] C. Baier and J.-P. Katoen. *Principles of Model Checking*. The MIT Press, 2008.

[2] E. Baykan, M. Henzinger, L. Marian, and I. Weber. A comprehensive study of features and algorithms for url-based topic classification. *TWEB*, 5(3):15, 2011.

[3] B. Berendt, L. Hollink, V. Hollink, M. Luczak-Rösch, K. H. Möller, and D. Vallet, editors. *USEWOD2012- 2nd International Workshop on Usage Analysis and The Web of Data*, Lecture Notes in Computer Science. Springer, 2012.

[4] R. E. Bucklin and C. Sismeiro. A Model of Web Site Browsing Behavior Estimated on Clickstream Data. *Journal of Marketing Research*, XL:249–267, Aug. 2003.

[5] L. D. Catledge and J. E. Pitkow. Characterizing browsing strategies in the world-wide web. In *Computer Networks and ISDN Systems*, pages 1065–1073, 1995.

[6] M. dÁquin, S. E. L., and E. Motta. Semantic technologies to support the user-centric analysis of activity data. In *Workshop on Social Data on the Web Workshop, SDoW 2011 at ISWC 2011*, 2011.

[7] D. Downey, S. T. Dumais, and E. Horvitz. Models of searching and browsing: Languages, studies, and application. In *Proceedings of IJCAI*, pages 2740–2747, 2007.

[8] I. Horrocks, O. Kutz, and U. Sattler. The even more irresistible $\mathcal{SROIQ}$. In *Proceedings of the 10th International Conference on Principles of Knowledge Representation and Reasoning (KR2006)*, pages 57–67, June 2006.

[9] J. Hoxha, M. Junghans, and S. Agarwal. Enabling semantic analysis of user browsing patterns in the web of data. In *USEWOD Workshop at the 21st International World Wide Web Conference (WWW2012)*, volume abs/1204.2713, 2012.

[10] T. Joachims, T. Hofmann, Y. Yue, and C.-N. J. Yu. Predicting structured objects with support vector machines. *Commun. ACM*, 52(11):97–104, 2009.

[11] E. J. Johnson, W. W. Moe, P. S. Fader, S. Bellman, and G. L. Lohse. On the depth and dynamics of online search behavior. *Manage. Sci.*, 50:299–308, March 2004.

[12] M.-Y. Kan and H. O. N. Thi. Fast webpage classification using url features. In *Proceedings of the 14th ACM international conference on Information and knowledge management*, CIKM '05, pages 325–326, New York, NY, USA, 2005. ACM.

[13] Y. LeCun, S. Chopra, M. Ranzato, and F. J. Huang. Energy-based models in document recognition and computer vision. In *ICDAR*, pages 337–341. IEEE Computer Society, 2007.

[14] N. R. Mabroukeh and C. I. Ezeife. Using domain ontology for semantic web usage mining and next page prediction. In *CIKM*, pages 1677–1680, 2009.

[15] A. Mehdi, P. Valtchev, R. Missaoui, and C. Djeraba. A framework for mining meaningful usage patterns within a semantically enhanced web portal. In B. C. Desai, C. K.-S. Leung, and S. P. Mudur, editors, *C3S2E*, ACM International Conference Proceeding Series, pages 138–147. ACM, 2010.

[16] A. L. Montgomery and C. Faloutsos. Identifying web browsing trends and patterns. *Computer*, 34:94–95, July 2001.

[17] D. Oberle, B. Berendt, A. Hotho, and J. Gonzalez. Conceptual user tracking. In E. M. Ruiz, J. Segovia, and P. S. Szczepaniak, editors, *AWIC*, volume 2663 of *Lecture Notes in Computer Science*, pages 155–164. Springer, 2003.

[18] Y. H. Park and P. S. Fader. Modeling browsing behavior at multiple websites. *Marketing Science*, pages 280–303, 2004.

[19] R. Srikant and R. Agrawal. Mining generalized association rules. In *Proceedings of the 21th International Conference on Very Large Data Bases*, VLDB '95, pages 407–419, San Francisco, CA, USA, 1995. Morgan Kaufmann Publishers Inc.

[20] R. Srikant and R. Agrawal. Mining sequential patterns: Generalizations and performance improvements. In P. M. G. Apers, M. Bouzeghoub, and G. Gardarin, editors, *EDBT*, volume 1057 of *Lecture Notes in Computer Science*, pages 3–17. Springer, 1996.

[21] R. Stühmer, D. Anicic, S. Sen, J. Ma, K.-U. Schmidt, and N. Stojanovic. Lifting events in rdf from interactions with annotated web pages. In *Proceedings of the 8th International Semantic Web Conference*, ISWC '09, pages 893–908, Berlin, Heidelberg, 2009. Springer-Verlag.

[22] G. Stumme, A. Hotho, and B. Berendt. Usage mining for and on the semantic web. In *Next Generation Data Mining. Proc. NSF Workshop*, pages 77–86, Baltimore, 2002.

[23] B. Taskar, C. Guestrin, and D. Koller. Max-margin markov networks. In S. Thrun, L. K. Saul, and B. Schölkopf, editors, *Advances in Neural Information Processing Systems 16 [Neural Information Processing Systems, NIPS 2003, December 8-13, 2003, Vancouver and Whistler, British Columbia, Canada]*. MIT Press, 2003.

[24] I. Tsochantaridis, T. Hofmann, T. Joachims, and Y. Altun. Support vector machine learning for interdependent and structured output spaces. In *Proceedings of the twenty-first international conference on Machine learning*, ICML '04, pages 104–, New York, NY, USA, 2004. ACM.

[25] M. Tvarožek, M. Barla, and M. Bieliková. Personalized presentation in web-based information systems. In *Proceedings of the 33rd conference on Current Trends in Theory and Practice of Computer Science*, SOFSEM '07, pages 796–807, Berlin, Heidelberg, 2007. Springer-Verlag.

[26] M. Vanzin, K. Becker, and D. D. A. Ruiz. Ontology-based filtering mechanisms for web usage patterns retrieval. In *EC-Web'05*, pages 267–277, 2005.

[27] H. Yilmaz and P. Senkul. Using ontology and sequence information for extracting behavior patterns from web navigation logs. In *Data Mining Workshops (ICDMW), 2010 IEEE International Conference on*, pages 549 –556, dec. 2010.