Quest: An Adaptive Framework for User Profile Acquisition from Social Communities of Interest

Nima Dokoohaki
Royal Institute of Technology (KTH)
Forum 120, 16440 Kista, Sweden
Email: nimad@kth.se

Mihhail Matskin
Norwegian University of Science and Technology (NTNU)
Trondheim, Norway
Email: misha@kth.se

***Abstract*—Within this paper we introduce a framework for semi- to full-automatic discovery and acquisition of bag-of-words style interest profiles from openly accessible Social Web communities. To do such, we construct a semantic taxonomy search tree from target domain (domain towards which we're acquiring profiles for), starting with generic concepts at root down to specific-level instances at leaves, then we utilize one of proposed Quest methods, namely Depth-based, N-Split and Greedy to read the concept labels from the tree and crawl the source Social Network for profiles containing corresponding topics. Cached profiles are then mined in a two-step approach, using a clusterer and a classifier to generate predictive model presenting weighted profiles, which are used later on by a semantic recommender to suggest and recommend the community members with the items of their similar interest.**

***Index Terms*—user profiles, profile mining, adaptive crawling, social network analysis.**

I. Introduction

Growth of Social Web, accompanied by much anticipation from Social Networking Web sites, has drawn a lot of attention from research community. As Web of interrelated content is gradually giving its place to Web of interpersonal content, classic problems of information retrieval continue to persist. At the same time publication means are becoming more and more easily available to both human and machine publishers. As these publication mediums increase their production pace day by day, it becomes harder for human readers to find and retrieve the exact content they are looking for. Adaptive Web and its myriads of approaches, specifically recommendation techniques and approaches have proven to be good candidates in dealing with retrieval of relevant information, by providing users with suggested contents of their taste. One infamous problem for recommendation techniques to function properly, is the sparsity of the usage data. One might want to build a recommender at the top of a content library already available, to provide users with suggestions while lack of enough user data hinders the functionality of the system. To deal with this problem, one can propose for discovery of interested users and acquisition and processing of their profiles for possibly generating recommendations of items according to their profiled interests. To do such we propose for a semi- to full-automatic framework, mainly composed of a crawler and a two step miner, in which crawler harvests the source repository for profiles and caches them. After normalization and dataset preparation, miner reads and analyzes raw profiles and applies clustering to gathered data to generate clusters of interrelated topics and their corresponding interests. Clusters are then stored and fed into classifier to create an initial probability distribution over all interrelated topic set, which could be used by a semantically enhanced recommender as initial point to generate recommendation for users of the system, seeking potential users or customers. For crawler to be capable of adaptively discovering related profiles, we build a taxonomy tree from the domain towards which we are discovering profiles for. And then we use the tree to formulate queries that are used eventually by crawler to cache the discovered profiles. To evaluate this framework, we have defined three mining schemes, namely: Depth-based, figure 3, focused crawling on topics on a certain tree-depth at each time, N-Split, iteratively focused crawling on all topics and N times splitting gathered data at each iteration, and Greedy, crawling the network for all topics and processing the cached data altogether, in a greedy fashion. We study this framework in the context of discovering possible interested on-line customers from LiveJournal [12], an openly accessible Social Blogging Network. We target the crawling process towards two startup on-line Museums, seeking to recommend their items to interested customers. We study the accuracy of the miner's clustering and classification performance with respect to each scheme proposed. The rest of this manuscript is structured as follows: first a brief background overview will be given, and then approach is introduced by studying the taxonomy tree structure, followed by introduction of three mining schemes proposed. This is followed by introduction of the framework, while last section describes the experiment with two Museums(Smartmuseums) followed by evaluation results. And finally a conclusion and future work section brings this manuscript to its end.

II. Related Work

User profiles play crucial role in the context of adaptivity enabled, and personalization empowered computer systems, their availability is vital for these systems to function properly. As a result, we see the problem from two perspectives: firstly, discovering the users and acquiring the knowledge pertaining to their profiles, and secondly, applying data mining techniques to these gathered data for inferring profiles which

can be consumed by a recommender for personalized recommendation generation. Mining profiles for personalization has been attractive to many fields such as collaborative filtering recommenders [1]. To deal with lack of user profiles, researchers approach different methodologies to gather, analyze and generate user profiles. For instance, in the field of personalization, techniques such as clustering of user transactions [4], data mining techniques such as clustering [2] and Web Usage Mining techniques, [6] were proposed. Agent-assisted personalization has been a target domain for applied user profiling. Soltysiak and Crabtree [7] give an overview of agent-based approaches to user profiling and introduce an learning approach to automatic generation of profiles. This learning cycle is made of a clusterer and a classifier with a feedback loop. The core of this learning process is a classical text classification approach. Sebastiani [24] gives an overview into machine learning heuristics for automated text categorization. In a similar work, Billsus and Pazzani [23] utilize the same heuristics for learning Web user profiles. We have adopted this mining scheme [7], [23] to learn and generate profiles in our framework. Most of the literature at hand are focused on generating aggregate usage history from an already cached and stored usage experience. Although not much attention has been payed to addressing the problem of formulating profiles when no experience data is available. That's why discovery of users has becomes crucial to assist the task of profile generation [8]. At the same time, if interests of the users are already at hand, we can utilize these interest descriptions for creation of possible user profiles. Users tend to often express these interest in the form of bag-of-words [13] or simply *topics*. As a result scientists are proposing new methodologies for generating profiles for users from these topics and utilizing them for generating recommendations [3]. At the same time, Social Web can serve as a novel source for discovery and acquisition of these topic-based interest profiles. Being seen as a data source, data mining approaches could be used to uncover patterns of profiled user data (e.g. interests) from Social Networks and services. Jensen and Neville [10] give an overview of statistical approaches applicable to a Social Network setting. Kleinberg [9] mentions approaches used for mining Social Networking sites. In this paper, we combine social network analysis and profile mining techniques, but we put more attention to the actual profile generation process. We focus our experiments on data acquired from Social blogging site, LiveJournal [12].

III. Taxonomy Assisted Crawling and Mining Social Communities of Interest

Many Social Networking sites, are paving the path for ease of publication and sharing of the content of the users within their boundaries. At the same time ease of publication and open expression of interests, allow for adaptive technologies to spot and attract customers to on-line libraries of materials or goods, which could be eventually purchased by interested customers [11]. We design a framework at the top of the source Social Network, which allows for a semi-automatic supervised approach utilizing a taxonomy assisted crawler and a two-step miner to first of all gather,then process and finally generate what we refer to as user profiles. Through the knowledge flow created across this framework, we harvest the crawled interest topics, process them through a two step learner, we digest them into clusters and classify them and eventually creates a predictive model of clusters along with their probabilistic weights (our generated profiles), that can be processed by semantic recommender for predictive recommendation generation. As designing the mining process has been the main focus of this work, we leave the recommendation generation for the future work. In the following section we focus on construction of the taxonomy tree, followed by the descriptions of crawling strategies and how they affect the overall knowledge flow.

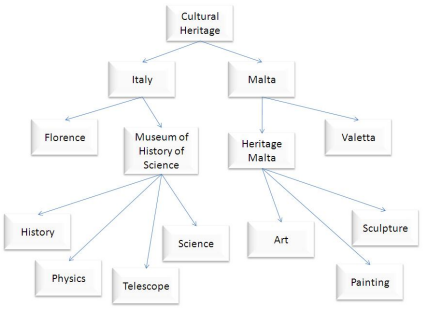

Fig. 1: *Adaptive taxonomy tree*

A. Constructing Adaptive Taxonomy from Target Domain

In order to specify those topics which are used for discovering and acquisition of relevant data for mining segment to work on, we should have a partial to rather complete knowledge of which topics are required to query for. This knowledge could be described semantically utilizing a taxonomical hierarchy comprising of topics surrounding the target domain, towards which we're gathering and processing data for. Taxonomy assisted, and later ontology-driven, adaptive crawling has been the subject of active research [14]. Ester et al., [15] introduce a framework for adaptive Web Crawling, which allows for interests of users to be expressed for the crawling purposes, and the framework introduces a schematic crawling algorithm for the purpose of the work. We have adopted this idea as the solution to the problem at hand. A taxonomy is defined as a tree consisting of nodes representing distinguished topics. In a general case taxonomy could be derived or constructed from an ontological presentation of the domain. In such a taxonomy edges represent the subtopic relationship. In our experiment, Figure1, topics are chosen from general topics corresponding to cultural heritage domain (root topic), and as we move to lower levels we see the branches towards each museum, while museum itself forms a concept and the leaves present the exhibits corresponding to the museums, so for the artistic museum we can observe topics related to art such as sculpture, while in the case of scientific exhibit, we observe topics such as telescope. It is worth mentioning that this taxonomical classification could be

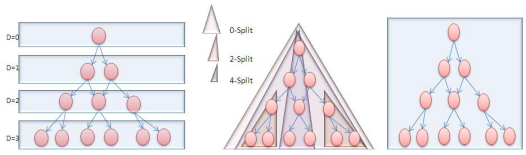

Fig. 2: *Schematics for Depth-based, N-Split and Greedy Quest*

more complete, and at the same time more complex.

B. Modeling and Designing Quest Strategies

To study the effectiveness of the approaches proposed for discovering and retrieval of possible matches across the network, we have proposed for a strategy set among which depending on the need for effectiveness or efficiency, strategy designer or mining supervisor can make a choice. If Φ is query set, then $\Phi = \{T_1, T_2, T_3, ..., T_n\}$ is a query that contains n topics, which is sent to interest repositories of Social Network to retrieve matching profiles, follows;

Depth-based is an iterative decremental strategy in which at each iteration, starting from the root, topics are taken from a certain depth of taxonomy tree structure ,*d* and are used to fill the query Φ to harvest the network for matching interest results. On the next iteration next depth *d+1* is taken into account, and the same approach is applied, and this mechanism iterates until the leaves of the taxonomy tree are reached. For example, in fig.2 taxonomy tree has depth d=4, and strategy iterates 4 times, accordingly. As a result:

$$\Phi_{depth} = \{\lceil T_1 \rceil_{d=0}, \lceil T_2, T_3, \rceil_{d=1}, \lceil T_{n-l}...T_n \rceil_d\}$$

Where Φ_{depth} is the query formulated according to Depth-based scheme, $\lceil T_{n-1}...T_n \rceil_d$ is a set of n topics taken from depth *d* of the taxonomy tree.

N-Split is an iterative strategy through which a subset (split) of the taxonomy tree is digested into query formula to be executed, at each iteration, until the size of split, becomes lower or equal to a single topic. In this strategy, it's important which number of splits are chosen. For instance, in fig.2, 3 splits are displayed as triangles of different size, in even splits. As a result:

$$\Phi_{split} = \{\lceil T_1...T_4 \rceil, \lceil T_4...T_8 \rceil, \lceil T_8...T_{12} \rceil, \lceil T_{n-4}...T_n \rceil_N\}$$

Where Φ_{split} is the query formulated according to Split-based scheme, $\lceil T_{n-4}...T_n \rceil$ is a split Topic set, divided by the size of N, e.g. N=4.

Greedy or all-at-once strategy, is a strategy, in which all topics within the tree are taken at once and are utilized to fill the query formula. As a result:

$$\Phi_{Greedy} = \{T_1...T_n\}$$

In which, Φ_{Greedy} is the query formulated according to Greedy strategy, and $\{T_1...T_n\}$ is a set on n topics that exist in the taxonomy tree.

C. Mining for Quest

While quest strategies allow for intelligent query formulation against the interest topics available in the profiles, retrieved matches need to be processed in the next step accordingly to generate profiles consumable by recommender system. To deal with interest topics gathered from the Social Network, we have proposed for a two-step mining process. In this process we reduce dimensionality of topic attributes through a clustering methodology, and we form centroid around the topics which are originally taken from taxonomy tree. Clusters are respectively fed into a function, typically a classifier, which eventually generates our initial (weighted) profiles.

1) Clustering Discovered Interest Topics : If T_i is taken as the topic that we have queried for, we define T'_i as the retrieved topic set corresponding to the latter query. As a matter of fact, T'_i can be taken into account as an observation set, $T'_i = \{T'_1, ..., T'_n\}$, where each observation can be seen as a d-dimensional vector, then a *k-means clustering* algorithm can be utilized to partition n observations into k sets $(K < n)$, in which $T'_{clustered} = \{T'_{clustered1}, ..., T'_{clusteredK}\}$ where $T'_{clustered}$ is the partitioned or clustered set. Interestingly, when adaptive taxonomy trees are built, we can observe that centroids of the clustering experiments are formed around those topics which are exactly matching topics to taxonomy tree, or are semantically close to those topics. We take advantage of this obvious effect to eliminate the centroids which are either irrelevant or less-relevant. With respect to this, the centroid formations are observed and supervised to make sure that the correct centroids are chosen for the clusters being formed. At the same time, number of clusters affect the centroid formation as well.

2) Classifying Clustered Interest Topics : In general case, clustering techniques are undertaken to reduce the dimensionality of input data, specially if number of input attributes are high. At the same time clustering allows a set of instances to be treated as a single (clustered) instance. As a result this can increase the effectiveness of the miner, specially if the output of clusterer is fed into input of a classifier. *Classification techniques* are very appealing where a predictive output based on a set of observations is needed. Since clustered topics are formed around the appealing topics, which correspond to taxonomy tree, we can use them as the input of a classifier. The classification output assigns higher predictive probabilities to more appealing topics, which are the topics we've queried for and now form the centroids of the clusters. As a result,

$$f'_{profiles} = f_{Classifier}(T'_{clustered})$$

Taken into account this fact, a classifier can be seen as a function, $f_{Classifier}$, where the set of our clustered observations $T'_{clustered}$, can be given to it as an input, while the output would be $f'_{profiles}$, a probabilistic model created from the clustered interest topics of respective users. At this point supervisor observes the predicted model and if needed refines the prediction output. Supervisor needs to observe the accuracy and precision of the experiments to make sure the classification step was successful. Predicted model

is stored for recommendation service to take as input and generate recommendation accordingly. Recommender system used here, is a semantic context-aware recommendation engine where recommendations are generated based upon their lexico-semantic distance from user interests to item annotations. Along with the topics, service also takes into account a weight value, which helps the process of recommendation query expansion and information retrieval precision. At each service call, a profile is retrieved along with their interest topics and probabilities generated for those clusters and are sent to engine for generating recommendations. For more information regarding recommendation process reader is advised to refer to Routsalo et al., [17].

IV. Evaluating Quest: a LiveJournal Experiment

A. Crawling LiveJournal Community Profiles

LiveJournal [12] is a Social Networking website empowering the users who share the passion of reading and writing. Users within this website can keep a journal, diary or simply a Blog [18]. LiveJournal has about 2 million active users per month. When users create profiles on this website, they emphasize the interests they have using a set of topics. LiveJournal website allows these interest topics [19] to be exported, XML formated in FOAF [20] vocabulary. There are in general two types of accounts on this website: *users* and *communities*. A LiveJournal community is a journal where users can share items, information and posts about a similar subject. Since communities are focused on a subject of interest, interest topics expressed for these communities creates this opportunity to study and analyze these communities for discovering interested customers, which turn out to be the members of the respective communities. We managed to discover, crawl and cache about 300 community profiles with in total 9750 topics, among which each profile contains at least one topic corresponding to existing topics in the taxonomy tree. Subsequently, we fed the data into the miner. First raw gathered topics are normalized using a filter. Then at each pace clusterer slices the data, while supervisor observes the clustering output and accuracy and then clustering results are saved and loaded into classifier to be processed for predicting the initial user profiles. At the end of this step supervisor observes the accuracy and resulting predictions before submitting it to recommender engine. Smartmuseum [21] is referred to a museum that runs Smartmuseum platform. This platform is a decentralized ontology-based knowledge management and dissemination system, capable of profiling [16] users and generating personalized recommendation [17] for them, which in this case are on-line or on-site visitors of website or physical exhibitions of museum, equipped with this software platform. Both museums subject to the experiment have created ontological descriptions of their items, for the platform to maintain and use. From these annotations we have built the taxonomy tree. At the moment this taxonomy hierarchy is built manually but for constructing this tree, an automated mechanism can be proposed based on ontology-learning schemes [22].

B. Evaluation and Comparison of Quest Strategies

In a controlled environment, we have implemented the framework to gather data from LiveJournal towards the Smartmuseum domain. As acting supervisor, museum curators were asked to supervise the overall mining results. In order to study the effectiveness and accuracy of the process, we can study different aspects of the system. Since our focus is more on evaluation of the Quest strategies, which enable adaptive discovery of communities as well as mining and transforming them into tangible data usable by system for possible usage, we focus the evaluational part of this manuscript onto effect of mining part to input gathered from each of the Quest strategies. Per each strategy curators ran the miner with the following configurations: Depth-based approach with depth configuration d=0,1,2 corresponding to 3 levels of the tree, N-Split approach with splits N=2,4,6 and Greedy approach with three configurations, two partial dataset (namely part1 and part2), and an altogether data containing all topics. To cluster the interest topics gathered from LiveJournal we have utilized *simple K-means clustering algorithms*. All of the experiments were done with *Euclidean distance*. *WCSSR (within cluster sum of squares root)*, is the measure used to measure the effectiveness of the clusterers. A good clustering result should be able to minimize WCSSR in general. To study the effectiveness and accuracy of the clusterer we have gathered and plotted *(WCSSR)* per each strategy. To demonstrate effectiveness of each strategy, we have tested different sizes of clusters, e.g. 2,4,6,8,16 and 32. Results are depicted in fig.3.

In the case of Greedy result-sets, where partial parts, present unstable results. Although All-at-once, which presents the all accumulated data for this experiment, shows average error in comparison. Still overall error for Greedy data is high since the quantity of the data is considerably high. Moving to N-split experiment, we observe that Split-based approach yields way less errors among all strategies. Finally observing Depth driven strategies, we can observe that error accumulated during clustering experiment is less than Greedy but still gradually higher than N-split technique. This gradual increase in error is visible starting at the root of the tree, e.g. Depth 0, moving towards the lower depth of the tree, e.g. Depth 2. we have experimented with two classifiers; a lazy classifier, *kNN (K-Nearest Neighbor)*, and a tree classifier, *PDT (Pruned Decision Tree)*. kNN classifies objects based on closest training examples. We have used Euclidean distance as the base metric of kNN. Decision Trees build and utilizes a decision tree as a predictive model which maps observations about an item to conclusions about the item's target value. To evaluate the accuracy of the classifier, accuracy and precision of the profiles generated are analyzed. For evaluating the classification step, we have used *F-Score (F-Measure)* with respect to two classifiers being tested. F-measure considers both the precision and the recall of the test to compute the score. In general F-measure is calculated as the harmonic mean of precision and recall where *precision* is the count of correct values divided by the count of all returned values and *recall* is the number

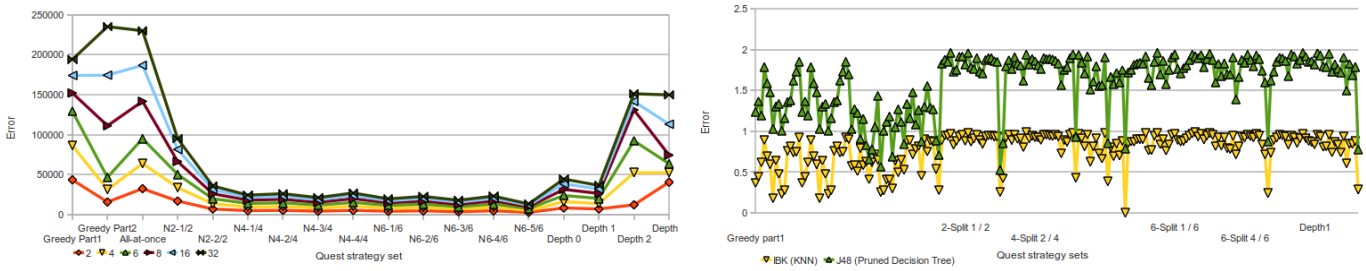

Fig. 3: Clustering (left) and Classification (right) accuracies for Quest strategies

of correct values divided by the number of values that should have been returned. As visualized in Fig.3, distribution and presentation of experimental data is the same as previous step for clustering. Although per each delimiter on horizontal axis, we have a single classified (predicted) cluster. kNN exhibits less error in comparison to PDT results. Starting with Greedy experiment set, F-Scores for partial sets is scattered, while in the case of All-at-once full data, error increases linearly per classified clusters. Interestingly in the latter case, error difference between PDT and kNN is not much, in comparison to formers. When moving to N-Split approach, we observe that f-scores with respect to both classifiers remain unchanged. What's interesting is that last splits are more accurate in each of the cases, e.g. N=2,4,6. It is worth mentioning that PDT and kNN exhibit the largest differences in F-scores in comparison to all other strategies. And finally while observing Depth-driven strategy we can observe slow increase in accuracy when moving down the root of the taxonomy tree, as depth increases.

V. Conclusion

In this work we presented a mining architecture which utilizes a set of adaptive strategies for effective discovery and mining user profiles utilized by a semantic recommender for recommendation generation. We experimented with interest profiles from a popular blogging Social Network and evaluated the performance of the two-step miner accordingly. Results present a trade-off between strategical approaches which could guide effective query formulation and gathering of profile data from Social Web for personalization and adaptive operations.

Acknowledgment

This work is partly supported by EU FP7 SMARTMUSEUM project, (FP7-216923) and grant number 621-2007-6565 funded by Swedish Research Council.

References

[1] Herlocker, J., Konstan, J., Borchers, A., and Riedl, J. 1999. An algorithmic framework for performing collaborative filtering. In Proceedings of the 1999 Conference on Research and Development in Information Retrieval, Aug. 1999.

[2] OConner, M. and Herlocker, J. 1999. Clustering items for collaborative filtering. In Proceedings of the ACM SIGIR Workshop on Recommender Systems, Berkeley, CA.

[3] Krestel, R., Fankhauser, P., and Nejdl, W. 2009. Latent dirichlet allocation for tag recommendation. In Proceedings of the Third ACM Conference on Recommender Systems, RecSys '09. ACM, New York, NY, 61-68.

[4] Mobasher, B. 1999. A web personalization engine based on user transaction clustering. In Proceedings of the 9th Workshop on Information Technologies and Systems (WITS99), Dec. 1999

[5] Buchner, A. and Mulvenna, M.D. 1999. Discovering internet marketing intelligence through on-line analytical web usage mining. SIGMOD Record, 4:27.

[6] Srivastava, J., Cooley, R., Deshpande, M., and Tan, P.-T. 2000. Web usage mining: Discovery and applications of usage patterns from Web data. SIGKDD Explorations, 1:2.

[7] Soltysiak SJ, Crabtree IB. Automatic Learning of User Profiles Towards the Personalisation of Agent Services. BT Technology Journal. 1998;16(3):110-117.

[8] Reichling T, Wulf V. Expert recommender systems in practice: evaluating semi-automatic profile generation. In: Boston, MA, USA: ACM; 2009:59-68.

[9] Kleinberg JM. Challenges in mining social network data: processes, privacy, and paradoxes. In: San Jose, California, USA: ACM; 2007:4-5.

[10] D. Jensen and J. Neville. Data mining in social networks. In National Academy of Sciences Symposium on Dynamic Social Network Modeling and Analysis, 2002.

[11] Banerjee, N., Chakraborty, D., Dasgupta, K., Mittal, S., Joshi, A., Nagar, S., Rai, A., and Madan, S. 2009. User interests in social media sites: an exploration with micro-blogs. CIKM '09. ACM, New York, NY, 1823-1826.

[12] LiveJournal. http://www.livejournal.com/

[13] Wallach, H. M. 2006. Topic modeling: beyond bag-of-words. In Proceedings of the 23rd international Conference on Machine Learning (Pittsburgh, Pennsylvania, June 25 - 29, 2006). ICML '06, vol. 148. ACM, New York, NY, 977-984.

[14] Novak B. A survey of focused web crawling algorithms. Proceedings of SIKDD. 2004:5558.

[15] Ester M, Kriegel H. Focused Web Crawling: A Generic Framework for Specifying the User Interest and for Adaptive Crawling Strategies. Computer.

[16] N. Dokoohaki and M. Matskin, "Personalizing human interaction through hybrid ontological profiling: Cultural heritage case study," in 1st Workshop on Semantic Web Applications and Human Aspects, (SWAHA08), M. Ronchetti, Ed., In conjunction with Asian Semantic Web Conference 2008. AIT e-press, 2008, pp. 133-140.

[17] T. Ruotsalo, E. Makela, T. Kauppinen, E. Hyvonen, K. Haav, V. Rantala, M. Frosterus, N. Dokoohaki, and M. Matskin, "Smartmuseum: Personalized context-aware access to digital cultural heritage," in Proceedings of the International Conferences on Digital Libraries and the Semantic Web 2009 (ICSD2009), Trento, Italy, 2009.

[18] A. Marwick . LiveJournal Users: Passionate, Prolific, and Private. LiveJournal, Inc. Research Report, New York, December 2008.

[19] LiveJournal Interests. http://www.livejournal.com/interests.bml

[20] FOAF Vocabulary Specification. http://xmlns.com/foaf/spec/.

[21] Smartmuseum platform. http://smartmuseum.eu/.

[22] Maedche, A. and Staab, S. 2001. Ontology Learning for the Semantic Web. IEEE Intelligent Systems 16, 2 (Mar. 2001), 72-79.

[23] Pazzani M, Billsus D. Learning and Revising User Profiles: The Identification of Interesting Web Sites. Machine Learning. 1997;27(3):313-331.

[24] Sebastiani F. Machine learning in automated text categorization. ACM Computing Surveys. 2002;34(1):1-47.