Collaborative Environments: Fine-Grained Knowledge Sharing- Review Paper

Mr. Naresh Patil [1], Prof. S. R. Durugkar[2]

1Department of Comp. Engg., SNDCOE & RC, Maharashtra
Head, Department Comp. Engg, SNDCOE & RC, Maharashtra

Abstract— *In this paper, we are propose a new approach for combining item-based collaborative filtering with case based reasoning to pursue personalized information filtering in a knowledge sharing context. Functionally, our personalized information filtering approach allows the use of recommendations by peers with similar interests and domain experts to guide the selection of information deemed relevant to an active user's profile. We are developing Collaborative Environments: Fine-Grained Knowledge Sharing a flexible collaboration platform that enables secure and focused information sharing across organizations. Collaborative Environments uses two key technologies developed at ISI to support a new concept of fine-grained semantically controlled information visibility. The Hands infrastructure provides a semantic network-based data model, search and filtering capabilities, distributed systems support and fine-grained control of resource visibility.*

Keywords— Advisor Search, Text Mining, Dirichlet Processes, Graphical Models.

I. INTRODUCTION

Many computer security problems can be essentially reduced to separating malicious from non-malicious activities. This is, for example, the case of spam filtering, intrusion detection, or the identification of fraudulent behaviour. But, in general, defining in a precise and computationally useful way what is harmless or what is offensive is often too complex. To overcome these difficulties, most solutions to such problems have traditionally adopted a machine-learning approach, notably through the use of classifiers to automatically derive models of (good and/or bad) behavior that are later used to recognize the occurrence of potentially dangerous events.

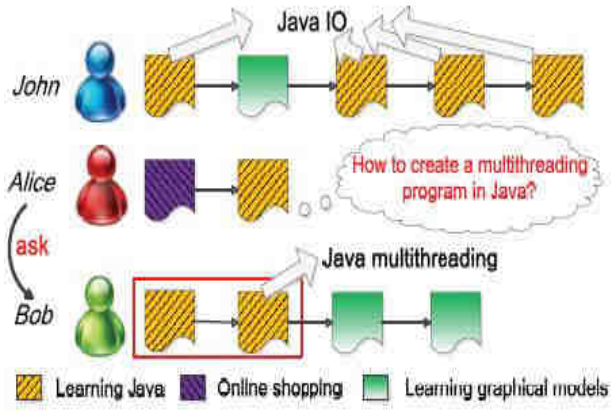

Fig. 1: Example for Knowledge Sharing In A Collaborative Environment[1]

An illustrative toy example is given in Figure 1. One can use “tcpdump” to intercept a sequence of Web surfing activities (IP packets) for each member. The scene is, Alice starts to surf the Web and wants to learn how to develop a Java multithreading program, which has already been studied by Bob (red rectangle). In this case, it might be a good idea to consult Bob, rather than studying by herself. We aim to provide such recommendations by analyzing surfing activities automatically. In this example, not necessarily Bob is an expert in every aspect of Java programming; however, due to his significant surfing activities in Java multithreading, it is reasonable to assume that he has gained enough knowledge in this area so that he can help Alice (in practice we could set a threshold on the amount of related surfing data to test significance). Even if Bob is still learning, he could share his experiences in learning and possibly suggest good learning materials to Alice, thus saving Alice’s effort and time.

Strictly speaking, KIDS’ idea of “learning with a secret” is not entirely new: Wang et al. introduced in Anagram, another payload-based anomaly detection system that addresses the evasion problem in quite a similar manner. We distinguish here between two broad classes of classifiers that use a key. In the first group, that we term randomized classifiers, the classifier is entirely public (or, equivalently, is trained with public information only).

However, in detection mode some parameters (the key) are randomly chosen every time an instance has to be classified, thus making uncertain for the attacker how the instance will be processed. Note that, in this case, the same instance will be processed differently every time if the key is randomly chosen. We emphasize that randomization can also be applied at training time, although it may only be sufficiently effective when used during testing, at least as far as evasion attacks are concerned. KIDS belong to a second group, that we call keyed classifiers. The practicality of various types of cryptanalytic attacks depends on many factors: Attacks based on few ciphertext are better than attacks that require many ciphertext, known plaintext attacks are better than chosen plaintext attacks, no adaptive attacks are better than adaptive attacks, single key attacks are better than related key attacks, etc. Since it is difficult to quantify the relative importance of all these factors in different scenarios, we usually concentrate on the total running time of the attack, which is a single well defined number.

II. EXISTING SYSTEM

The major problem of computing optimal strategies to modify an attack so that it evades detection by a Bayes classifier. The problem can be formulated in game theoretic terms, where each modification made to an instance comes at a price, and successful detection and evasion have measurable utilities to the classifier and the adversary, respectively. The authors study how to detect such optimally modified instances by adapting the decision surface of the classifier, and also discuss how the adversary might react to this. The setting used in assumes an adversary with full knowledge of the classifier to be evaded. Shortly after, how evasion can be done when such information is unavailable. They formulate the adversarial classifier reverse engineering problem (ACRE) as the task of learning sufficient information about a classifier to construct attacks, instead of looking for optimal strategies. The authors use a membership oracle as implicit adversarial model: the attacker is given the opportunity to query the classifier with any chosen instance to determine whether it is labelled as malicious or not. Consequently, a reasonable objective is to find in-stances that evade detection with an affordable number of queries. ACRE learnable if there exists an algorithm that finds a minimal-cost in-stance evading detection using only polynomial many queries. Similarly, a classifier is ACRE k-learnable if the cost is not minimal but bounded by k. Among the results given, it is proved that linear classifiers with continuous features are ACRE k-learnable under linear cost functions. Therefore, these classifiers should not be used in adversarial environments. Subsequent work by generalizes these results to convex-inducing classifiers, showing that it is generally not necessary to reverse engineer the decision boundary to construct undetected instances of near-minimal cost. For the some open problems and challenges related to the classifier evasion problem. Additional works have revisited the role of ma-chine learning in security applications, with particular emphasis on anomaly detection.

III. SURVEY REVIEW

1. Matthew J. Beal show that it is possible to extend hidden Markov models to have a count ably infinite number of hidden states. By using the theory of Dirichlet processes we can implicitly integrate out the infinitely many transition parameters, leaving only three hyper parameters which can be learned from data. These three hyper parameters define a hierarchical Dirichlet process capable of capturing a rich set of transition dynamics. The three hyper parameters control the time scale of the dynamics, the sparsity of the underlying state-transition matrix, and the expected number of distinct hidden states in a finite sequence.
2. Krisztian Balog present two general strategies to expert searching given a document collection which are formalized using generative probabilistic models. The first of these directly models an expert's knowledge based on the documents that they are associated with, whilst the second locates documents on topic, and then finds the associated expert. Forming reliable associations is crucial to the performance of expert finding systems. Consequently, in our evaluation we compare the different approaches, exploring a variety of associations along with other operational parameters.
3. Hongbo Deng present three models for expert finding based on the large-scale DBLP bibliography and Google Scholar for data supplementation. The first, a novel weighted language model, models an expert candidate based on the relevance and importance of associated documents by introducing a document prior probability, and achieves much better results than the basic language model. The second, a topic-based model, represents each candidate as a weighted sum of multiple topics, whilst the third, a hybrid model, combines the language model and the topic-based model. We evaluate our system using a benchmark dataset based on human relevance judgments of how well the expertise of proposed experts matches a query topic.

4. Anil K. Jain provide a brief overview of clustering, summarize well known clustering methods, discuss the major challenges and key issues in designing clustering algorithms, and point out some of the emerging and useful research directions, including semi-supervised clustering, ensemble clustering, simultaneous feature selection during data clustering, and large scale data clustering..
5. Jie Bao motivates the need for collaborative environments for ontology construction, sharing, and usage; identifies the desiderata of such environments; and proposes package based description logics (P-DL) that extend classic description logic (DL) based ontology languages to support modularity and (selective) knowledge hiding. In P-DL, each ontology consists of packages (or modules) with well-defined interfaces. Each package encapsulates a closely related set of terms and relations between terms.

IV. KNOWLEDGE BASES, HAVE SEVERAL IMPORTANT CHARACTERISTICS

1. Constructing large ontologies typically requires collaboration among multiple individuals or groups with expertise in specific areas, with each participant contributing only a part of the ontology. Therefore, instead of a single, centralized ontology, in most domains, there are multiple distributed ontologies covering parts of the domain.

2. Because no single ontology can meet the needs of all users under every conceivable scenario, the ontology that meets the needs of a user or a group of users' needs to be assembled from several independently developed ontology modules. Since different ontologies or different modules of a single ontology are developed by people with diverse points of view, semantic inconsistencies or conflicts between such modules are inevitable. Consequently, in collaborative ontology environments, there is a need for mechanisms for resolving or managing such semantic conflicts to ensure that the resulting ontology is not internally inconsistent.

3. While ontologies are often used to facilitate sharing of knowledge, data, and resources, many real-world scenarios also call for selectively hiding certain parts of an ontology (or conversely, selectively sharing certain parts of an ontology). The need for knowledge hiding may arise due to privacy and security concerns, or for managing and knowledge engineering purposes.

V. KNOWLEDGE HIDING IN ONTOLOGY

In this Fikes et al. mentioned integration of modular ontology's in the Ontolingua system and restricting symbol access to public or private. The major difference between our approach and their approach is that we use packages not only as modular ontology units, but also in organizational hierarchies, therefore enabling the hierarchical management of modules in collaborative ontology building. The scope limitation modifier idea is an extension of the idea of symbol access restriction, but it is more flexible and expressive.

Efforts aimed at developing formal languages to control ontology access scope include Extensible Access Control Markup Language (XACML) and policy languages. Giereth studied hiding part of RDF, where sensitive data in an RDF-graph is encrypted for a set of recipients, while all non-sensitive data remain publicly readable. However, those efforts are aimed at safe access on language or syntactic level. On the other hand, SLM in P-DL aims at knowledge hiding on semantic level, where the hiding is not total, but partial, i.e., hiding semantics can still be used in safe indirect inferences. Farkas studied unwanted inferences problem in semantic web data on XML, RDF or OWL level. Our approach to SLM and concealable reasoning is a more principled formalism to avoid unwanted inferences and with better defined localized semantics.

VI. EXPERT SEARCH

Expert search aims at retrieving people who have expertise on the given query topic. Early approaches involve building a knowledge base which contains the descriptions of people's skills within an organization. Expert search became a hot research area since the start of the TREC enterprise track in 2005. Balog et al. proposed a language model framework for expert search. Their Model 2 is a document centric approach which first computes the relevance of documents to a query and then accumulates for each candidate the relevance scores of the documents that are associated with the candidate. This process was formulated in a generative probabilistic model. Balog et al. showed that Model 2 performed better and it became one of the most prominent methods for expert search. Other methods have been proposed for enterprise expert search, but the nature of these methods is still accumulating relevance scores of associated documents to candidates. Expert retrieval in other scenarios has also been studied, e.g. online question answering communities academic society.

CONCLUSION

Finally we conclude that our study of fine-grained knowledge sharing in collaborative environments, which identified digging out fine grained knowledge reflected by people's interactions with the outside world as the key to solving this problem. System proposes a two-step

framework to mine fine-grained knowledge and integrated it with the classic expert search method for finding right advisors. We demonstrate the feasibility of mining task micro-aspects for solving this knowledge sharing problem. We leave these possible improvements to future work. Finally, the classic expert search method is applied to the mined results to find proper members for knowledge sharing. Experiments on Web surfing data collected from our lab at UCSB and IBM show that the fine grained aspect mining framework works as expected and outperforms baselines.

REFERENCES

[1] Ziyu Guan, Shengqi Yang, Huan Sun, "Fine-Grained Knowledge Sharing in Collaborative Environments", IEEE Transactions on Knowledge and Data Engineering.

[2] K. Balog, L. Azzopardi, and M. de Rijke. Formal models for expert finding in enterprise corpora. In SIGIR, pages 43–50, 2006.

[3] M. J. Beal, Z. Ghahramani, and C. E. Rasmussen. The infinite hidden markov model. In Advances in neural information processing systems, pages 577–584, 2001.

[4] M. Belkin and P. Niyogi. Laplacian eigenmaps and spectral techniques for embedding and clustering. In NIPS, 2001. 25] Y. Teh, M. Jordan, M. Beal, and D. Blei. Hierarchical dirichlet processes. Journal of the American Statistical Association, 101(476):1566–1581, 2006.

[5] J. Van Gael, Y. Saatci, Y. Teh, and Z. Ghahramani. Beam sampling for the infinite hidden markov model. In ICML, pages 1088–1095, 2008.

[6] U. Von Luxburg. A tutorial on spectral clustering. Statistics and computing, 17(4):395–416, 2007.

[7] H. Wang, Y. Song, M.-W. Chang, X. He, R. White, and W. Chu. Learning to extract cross-session search tasks. In WWW, pages 1353–1364, 2013.

[8] R. White, P. Bailey, and L. Chen. Predicting user interests from contextual information. In SIGIR, pages 363–370, 2009.

[9] Y. Zhao, G. Karypis, and U. Fayyad. Hierarchical clustering algorithms for document datasets.

[10] D. Blei and M. Jordan. Variational inference for dirichlet process mixtures. Bayesian Analysis, 1(1):121–143, 2006.

[11] D. M. Blei, T. L. Griffiths, M. I. Jordan, and J. B. Tenenbaum. Hierarchical topic models and the nested chinese restaurant process. In NIPS, 2003.

[12] D. M. Blei and J. D. Lafferty. Dynamic topic models. In ICML, pages 113–120, 2006.

[13] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. Journal of machine Learning research, 3:993–1022, 2003.

[14] Y. Low, D. Bickson, J. Gonzalez, C. Guestrin, A. Kyrola, and J. M. Hellerstein. Distributed graphlab: a framework for machine learning and data mining in the cloud. In Proceedings of the 38th international conference on very large databases, pages 716–727, 2012.

[15] M. Medvedovic and S. Sivaganesan. Bayesian infinite mixture model based clustering of gene expression profiles. Bioinfor- matics, 18(9):1194–1206, 2002.

[16] R. M. Neal. Slice sampling. Annals of statistics, pages 705–741, 2003.

[17] C. Rasmussen. The infinite gaussian mixture model. In NIPS, page 554C560, 2000.