*The Knowledge Engineering Review*, Vol. 25:3, 359–364. 
doi:10.1017/S0269888910000226

# Abstracts of Recent PhDs

## Abstraction, Aggregation and Recursion for Generating Accurate and Simple Classifiers

**Candidate:** Dae-Ki Kang
**Institution:** Department of Computer Science, Iowa State University, Iowa, USA
**Supervisor:** Vasant Honavar
**Year awarded:** 2006
**URL:** http://www.cs.iastate.edu/~dkkang/

### Abstract

An important goal of inductive learning is to generate accurate and compact classifiers from data. In a typical inductive learning scenario, instances in a data set are simply represented as ordered tuples of attribute values. In our research, we explore three methodologies to improve the accuracy and compactness of the classifiers: abstraction, aggregation, and recursion.

First, abstraction is aimed at the design and analysis of algorithms that generate and deal with taxonomies for the construction of compact and robust classifiers. In many applications of the data-driven knowledge discovery process, taxonomies have been shown to be useful in constructing compact, robust, and comprehensible classifiers. However, in many application domains, human-designed taxonomies are unavailable. We introduce algorithms for automated construction of taxonomies inductively from both structured (such as UCI Repository) and unstructured (such as text and biological sequences) data. We introduce AVT-Learner, an algorithm for automated construction of attribute value taxonomies (AVT) from data, and Word Taxonomy Learner (WTL), an algorithm for automated construction of word taxonomy from text and sequence data. We describe experiments on the UCI data sets and compare the performance of AVT-NBL (an AVT-guided Naive Bayes Learner) with that of the standard Naive Bayes Learner (NBL). Our results show that the AVTs generated by AVT-Learner are competitive with human-generated AVTs (in cases where such AVTs are available). AVT-NBL using AVTs generated by AVT-Learner achieves classification accuracies that are comparable to or higher than those obtained by NBL; and the resulting classifiers are significantly more compact than those generated by NBL. Similarly, our experimental results of WTL and WTNBL on protein localization sequences and Reuters newswire text categorization data sets show that the proposed algorithms can generate Naive Bayes classifiers that are more compact and often more accurate than those produced by the standard Naive Bayes learner for the Multinomial Model.

Secondly, we apply aggregation to construct features as a multi-set of values for the intrusion detection task. For this task, we propose a bag of system calls representation for system call traces and describe the misuse and anomaly detection results on the University of New Mexico (UNM) and MIT Lincoln Lab (MIT LL) system call sequences with the proposed representation. With the feature representation as input, we compare the performance of several machine-learning techniques for misuse detection and show experimental results on anomaly detection. The results show that standard machine-learning and clustering techniques using the simple bag of system calls representation based on the system call traces generated by the operating system's kernel is effective and often performs better than approaches that use foreign contiguous sequences in detecting intrusive behaviors of compromised processes.

Finally, we construct a set of classifiers by recursive application of the Naive Bayes learning algorithms. The Naive Bayes (NB) classifier relies on the assumption that the instances in each class can be described by a single generative model. This assumption can be restrictive in many real-world classification tasks. We describe the recursive Naive Bayes learner (RNBL), which relaxes this assumption by constructing a tree of Naive Bayes classifiers for sequence classification, where each individual NB classifier in the tree is based on an event model (one model for each class at each node in the tree). In our experiments on protein sequences, Reuters' newswire documents and UC-Irvine benchmark data sets, we observe that RNBL substantially outperforms the NB classifier. Furthermore, our experiments on the protein sequences and the text documents show that RNBL outperforms the C4.5 decision tree learner (using tests on sequence composition statistics as the splitting criterion) and yields accuracies that are comparable to those of support vector machines (SVM) using similar information.

---

## Ontology-based Discovery and Composition of Geographic Information Services

**Candidate:** Michael Lutz
**Institution:** Institute for Geoinformatics, University of Muenster, Germany
**Supervisors:** Werner Kuhn and Ubbo Visser
**Year awarded:** 2006
**URL:** http://ifgi.uni-muenster.de/~lutzm/

## Abstract

Spatial data infrastructures will greatly benefit from the ability to compose geographic information (GI) services to solve complex problems. Discovering suitable services for data access and geoprocessing is a major challenge in this endeavour. Current (keyword-based) approaches to service discovery are inherently restricted by the ambiguities of the natural language, which can lead to low precision and/or recall. To alleviate these problems, we propose two ontology-based approaches for enhanced discovery of GI services. The approach for ontology-based discovery of data access services is based on semantic matchmaking between Description Logic (DL) concepts representing geographic feature types and the requester's query. DL subsumption reasoning is used to find matches between queries and service descriptions. The approach for ontology-based discovery of geoprocessing services rests on two ideas.

Ontologies describing geospatial operations are used to create descriptions of user requirements and service capabilities. Matches between these descriptions are identified based on function subtyping. In both approaches, service descriptions are based on a shared vocabulary that contains the basic terms of a domain and for which a shared understanding between the actors in the domain is assumed. We use a running example from the geospatial domain to analyze those problems that can occur in the existing keyword- and ontology-based approaches and how the discovery of GI services differs from other service discovery tasks. The example is also used for illustrating the prototypical implementation of the proposed approach.


---

## Cognitive Agent Programming: A Semantic Approach


**Candidate:** M. Birna van Riemsdijk
**Institution:** Department of Information and Computing Sciences, Utrecht University, The Netherlands
**Supervisors:** John-Jules Ch. Meyer, Frank S. de Boer and Mehdi Dastani


**Year awarded:** 2006
**URL:** http://www.pst.ifi.lmu.de/people/staff/riemsdijk



## Abstract

In this thesis, we are concerned with the design and investigation of dedicated programming languages for programming agents. We focus in particular on programming languages for rational agents, that is, flexibly behaving computing entities that are able to make 'good' decisions about what to do. An important line of research in this area is based on Bratman's so-called Belief Desire Intention (BDI) philosophy. The idea of BDI philosophy is that the behavior of rational agents can be predicted by ascribing beliefs, desires, and intentions to the agent, and by assuming that the agent will tend to act in pursuit of its desires, taking into account its beliefs about the world. The idea was then coined at the beginning of the 1990s that it might not only be possible to explain and describe rational agents in terms of the BDI notions, but that it might also be possible to program rational agents, using these notions as first class citizens in a programming language. The research that is done along these lines not only uses the notions of beliefs, desires, and intentions, but also related notions such as goals and plans. We refer to these notions as 'cognitive' notions, and to programming languages for agents based on these notions as *cognitive agent programming languages*. Our work proposes new constructs for representing these cognitive notions in a programming language and investigates existing constructs. We take a semantic approach, in that we define formal semantics for the proposed constructs and investigate the constructs by performing a semantic analysis. We investigate, in particular, ways for representing goals, and we study a construct called *plan revision rule* of the cognitive agent programming language, 3APL, which can be used for revising an agent's plan if the circumstances call for this.


---

## Goal-Directed Complete-Web Recommendation


**Candidate:** Tingshao Zhu
**Institution:** Department of Computing Science, University of Alberta, Alberta, Canada
**Supervisors:** Russ Greiner and Gerald Haeubl


**Year awarded:** 2006
**URL:** http://www.cs.ualberta.ca/ ~ tszhu



## Abstract

While the World Wide Web (WWW) contains a vast quantity of information, it is often difficult for web users to find the information they seek. Here, a passive Goal-Directed Complete-Web (GCW) recommender system, which recommends relevant pages from anywhere on the web to satisfy the user's current information need without any explicit additional input, has been developed. After identifying the search strategy that is employed by actual users while they browse the web, the model attempts to locate the pages that satisfy the user's information need based on the content of the pages the user has visited, and the actions the user has applied to these pages. To build such models, I develop a number of browsing features—browsing properties of the words, in the context of the current session—to capture the actions of the web user. Because the method is based on how the words are used (while training on these browsing feature values), it can be applied to make predictions about pages that have never been visited. This model is therefore independent of users, specific words and specific web pages, and so it can be used to identify relevant pages in any new web

environment. To evaluate the predictive models, we have conducted two user studies, each involving over one hundred participants. Data from the user studies demonstrate that the models can effectively identify the information needs of new users, leading them to previously unseen, but relevant pages.

## Representing and Reasoning with Modular Ontologies

**Candidate:** Jie Bao
**Institution:** Department of Computer Science, Iowa State University, Iowa, USA
**Supervisors:** Vasant Honavar
**Year awarded:** 2007
**URL:** http://www.cs.iastate.edu/ baojie/acad/JieBaoDissertation.pdf

### Abstract

The success of the web can be partially attributed to the network effect: the web is realized by interlinked web pages contributed by independent actors. We expect the network effect will also play an important role in realizing the full potential of the next generation of web—the semantic web: instead of a single, centralized ontology, it is much more natural to have multiple, interlinked, and distributed ontologies. Such ontologies represent the contextual, local knowledge of the ontology designers. Ideally, ontology languages will support localized and contextualized semantics, partial and selective reuse of ontology modules, and collaborative construction of large ontologies. To address these issues, this dissertation develops a novel approach, namely Package-based Description Logics (P-DL). In particular, the following problems are identified and investigated:

- The identification and theoretical characterization of the desiderata of modular ontology languages that can support selective sharing and reuse of knowledge across independently developed knowledge bases;
- The development of the language P-DL, which extends the classical description logics (DL) to support selective knowledge sharing through a semantic importing mechanism and the establishment of a minimal set of restrictions on the use of imported symbols to support contextualized semantics and transitive propagation of imported knowledge.
- The development of a family of sound and complete tableau-based, federated reasoning algorithms for distributed P-DL ontologies including ALCP and SHIQP, thus the usually costly ontology integration can be avoided.
- The formulation of the criteria for answering queries against a knowledge base using hidden or private knowledge, whenever it is feasible to do so without compromising hidden knowledge, and the development of privacy-preserving reasoning strategies for the case of the commonly used hierarchical ontologies and SHIQ ontologies.
- The development of some prototype tools for collaborative development of large ontologies, including support for concurrent editing and partial editing of an ontology.

## Reasoning with Dynamic Networks in Practice

**Candidate:** Theodoros Charitos
**Institution:** Department of Information and Computing Sciences, University of Utrecht, the Netherlands
**Supervisor:** Linda C. van der Gaag
**Year awarded:** 2007
**URL:** http://www.cs.uu.nl/staff/theodore.html

### Abstract

The modelling and analysis of time-evolving phenomena constitute a significant topic in science and engineering. Dynamic Bayesian networks are powerful and flexible graphical models for representing and computing probability distribution over variables that relate to stochastic processes. For a set of variables capturing phenomena that evolve over time, such models specify a set of conditional independence assumptions that allow the joint distribution to be represented in a factored way. Efficient inference with such models, however, remains a crucial issue for their successful application in practice. The main contribution of this thesis lies in the methods for easing inference with dynamic networks by the exploitation of either the nature of the data or the parameters of the model. To provide for a realistic setting, the thesis first considers the extension of a static model for the management of patients in intensive care units suspected of suffering from ventilator-associated pneumonia, into a dynamic network. Using this dynamic network and the real-life data accompanying it as the main vehicle upon which all methods are constructed, this thesis demonstrates the effect of the model's parameters on its output probability distribution or on a decision for antibiotic treatment based upon this distribution. The precise form of mathematical functions describing these effects is established, while approximate methods for the efficient computing of these functions are also proposed. Next, the thesis presents flexible inference algorithms that are tailored to the application at hand and exploit either consecutive similar values from diagnostic tests or symptoms of several patients that are sequentially observed, or the specifications of the probabilities in the transition matrices and in the sensitivity and specificity rates of several diagnostic tests of the model. Finally, the thesis presents an algorithm for efficient inference in the case in which the observed data arrive at arbitrary points in time instead of with pre-defined transition intervals. As a result, the distribution of the hidden variables is approximated at such time by interpolating between the boundaries of the pre-defined intervals while using the algorithm for further computations.

## Learning from Partially Labeled Data: Unsupervised and Semi-supervised Learning on Graphs and Learning with Distribution Shifting

**Candidate:** Jiayuan Huang
**Institution:** School of Computer Science, University of Waterloo, Ontario, Canada
**Supervisor:** Dale Schuurmans
**Year awarded:** 2007
**URL:** http://uwspace.uwaterloo.ca/bitstream/10012/3165/1/Thesis Jiayuan.pdf

### Abstract

This thesis focuses on two fundamental machine learning problems: unsupervised learning, where no label information is available, and semi-supervised learning, where a small number of labels is given in addition to unlabeled data. These problems arise in many real-world applications where a large amount of data is available, but no or only a small amount of labeled data exists. Obtaining classification labels in these domains is usually quite difficult because it involves either manual labeling or physical experimentation. This thesis approaches these problems from two perspectives: graph based and distribution based.

First, I investigate a series of graph-based learning algorithms that are able to exploit the information embedded in different types of graph structures. These algorithms allow label information to be shared between nodes in the graph—ultimately communicating information globally to yield effective unsupervised and semi-supervised learning. In particular, I extend the existing graph-based learning algorithms, currently based on undirected graphs to directed graphs, hypergraphs and complex networks. These richer graph representations allow one to more naturally capture the intrinsic data relationships that exist, for example, in web data, relational data, bioinformatics and social networks.

Second, I investigate a more statistically oriented approach that explicitly models a learning scenario where the training and test examples come from different distributions. This is a difficult situation for standard statistical learning approaches, since they typically incorporate an assumption that the distributions for training and test sets are similar, if not identical. I utilize unlabeled data to correct the bias between the training and test distributions. A key idea is to produce resampling weights for bias correction by working directly in a feature space and bypassing the problem of explicit density estimation. The technique can be easily applied to many different supervised learning algorithms, automatically adapting their behavior to cope with distribution shifting between training and test data.

---

## A Semantic-Aware Framework for Personalized Learning Objects Retrieval and Recommendation

**Candidate:** Ming Che Lee
**Institution:** Department of Engineering Science, National Cheng Kung University, Taiwan, Republic of China
**Supervisor:** Tzone I. Wang
**Year awarded:** 2007
**URL:** http://www.es.ncku.edu.tw

### Abstract

In recent years, distance learning has become more and more realistic and popular owing to the rapid advance in the Internet, especially in web–page interaction technology. Solving the problems of sharing and reusing teaching materials in different e-learning systems has been an important issue. There have been numerous international organizations devoted to building e-learning standards, such as IEEE-LTSC, Instruction Management System (IMS) Global Learning Consortium, Aviation Industry CBT Committee, Advanced Distributed Learning initiative, and also Alliance of Remote Instructional Authoring and Distribution Networks for Europe project. For the time being, the Sharable Content Object Reference Model (SCORM) is recognized as the most popular standard. It aims to foster the creation of reusable learning content used as 'instructional objects' within a common technical framework for both computer and web-based learning.

Learning Object Metadata (LOM) is approved by the IEEE–Standards Association. LOM aims to provide structured descriptions of digital contents, sometimes referred to as 'Learning Objects' (LOs). IEEE-LOM uses a pre-defined and common vocabulary to describe the content of learning objects. LOM plays the same role as the Dewey Decimal Classification (DDC) in library books catalog, which can guide a learner to locate learning objects on the Internet by using titles, descriptions, locations, and other attributes. However, in an e-learning environment a learner may not have any basic knowledge of a specific domain before learning it. The query terms input to a tutoring system for specific subject may only be a learner's surmises and the learning objects retrieved accordingly may be incorrect, worse, or misleading.

This thesis proposes an ontology-based framework for establishing personalized learning objects retrieval. In the proposed framework, Domain Ontology is used for constructing automatic inferring of user intention. The personalization functionality is provided by the probabilistic semantic inferring of user intention and user preference. An ontology query expansion algorithm and an integrated learning objects recommendation algorithm are proposed. Focused on digital learning material and contrasted with other traditional keyword-based search technologies, the proposed approach has shown significant improvement in retrieval recall precision rate, and recommendation performance.

## Interactive and Verifiable Web Services Composition, Specification Reformulation and Substitution

**Candidate:** Jyotishman Pathak
**Institution:** Department of Computer Science, Iowa State University, Iowa, USA
**Supervisor:** Vasant Honavar
**Year awarded:** 2007
**URL:** http://www.cs.iastate.edu/~jpathak

### Abstract

Recent advances in networks, information and computation grids, and WWW have resulted in the proliferation of a multitude of software components and services. These developments allow us to rapidly build new applications from existing ones. Toward this end, this dissertation develops solutions for the following problems related to Web services:

1. Web Service Composition: We propose a new framework for modeling complex web services based on the techniques of abstraction, composition and reformulation. The approach allows service developers to specify an abstract and possibly incomplete specification of the composite (goal) service. This specification is used to select a set of suitable component services such that their composition realizes the desired goal. In the event that such a composition is unrealizable, the cause(s) for the failure of composition is determined and communicated to the developer, thereby enabling further reformulation of the goal specification.
2. Web Service Specification Reformulation: At present, handling failure of composition requires the service developers to manually analyze the cause(s) of failure and model alternate goal specifications. To assist the developers in such situations, we describe a technique which given the specification of a desired composite service with a certain functional behavior, automatically identifies alternate specifications with the same functional behavior.
3. Web Service Substitution: We introduce the notion of context-specific substitutability in web services, where context refers to the overall functionality of the composition that is required to be maintained after the replacement of its constituents. Using the context information, we investigate two variants of the substitution problem, namely environment-independent and environment-dependent, where environment refers to the constituents of a composition and shows how the substitutability criteria can be relaxed within this model.

The work described above contributed to the design and implementation of MoSCoE, an open-source platform for modeling and executing complex web services.

---

## Activity Recognition for Agent Teams

**Candidate:** Gita Sukthankar
**Institution:** Robotics Institute, Carnegie Mellon University, PA, USA
**Supervisor:** Katia Sycara
**Year awarded:** 2007
**URL:** http://www.ri.cmu.edu/pubs/pub 5819.html

### Abstract

Proficient teams can accomplish the goals that would not otherwise be achievable by groups of uncoordinated individuals. This thesis addresses the problem of analyzing team activities from external observations and prior knowledge of the team's behavior patterns. There are three general classes of recognition cues that are potentially valuable for team activity/plan recognition: (1) spatial relationships between team members and/or physical landmarks that stay fixed over a period of time; (2) temporal dependencies between behaviors in a plan or between actions in a behavior; (3) coordination constraints between agents and the actions that they are performing. This thesis examines how to leverage available spatial, temporal, and coordination cues to perform off-line multi-agent activity/plan recognition for teams with dynamic membership.

In physical domains (military, athletic, or robotic), team behaviors often have an observable spatio-temporal structure, defined by the relative physical positions of team members and their relation to static landmarks. We suggest that this structure, along with the temporal dependencies and coordination constraints defined by a team plan library, can be exploited to perform behavior recognition on traces of agent activity over time, even in the presence of uninvolved agents. Unlike prior work in team plan recognition, where it is assumed that team membership stays constant over time, this thesis addresses the novel problem of recovering agent-to-team assignment for team tasks where team composition, the mapping of agents into teams, changes over time; this allows the analysis of more complicated tasks in which agents must periodically divide into sub-teams.

This thesis makes four main contributions: (1) an efficient and robust technique for formation identification based on spatial relationships; (2) a new algorithm for simultaneously determining team membership and performing behavior recognition on spatio-temporal traces with dynamic team membership; (3) a general pruning technique based on coordination cues that improve the efficiency of plan recognition for dynamic teams; and (4) methods for identifying player policies in team games that lack strong spatial, temporal, and coordination dependencies.

## Object Extraction in a Soft Computing Framework

**Candidate:** Siddhartha Bhattacharyya
**Institution:** Department of Computer Science and Engineering, Jadavpur University, Kolkata, West Bengal, India
**Supervisor(s):** Ujjwal Maulik
**Year awarded:** 2008
**URL:** http://www.jadavpur.edu

### Abstract

Conventional neural network architectures suffer from their inability to handle multidimensional image data, bipolar transfer characteristics and heavily interconnected topologies. In this thesis, an effort has been made to overcome the aforementioned shortcomings, with special reference to the neighborhood topology-based multilayer, self-organizing neural network (MLSONN) architecture. The initial steps in this direction are centered on the evolving, multilevel transfer characteristics for the network architecture for inducing multiscaling capabilities in the network architectures. A proposed multilevel MUSIG transfer function enables the network architecture to handle gray-scale/multilevel image data. A parallel extension of the conventional MLSONN architecture has been proposed for the extraction and segmentation of color images. The performance of the extended parallel version has been reported with several multilevel activation functions. In order to reduce the space and time complexity of the extraction procedures of the conventional MLSONN architecture, methods have been devised to refine the network interconnections by means of fuzzy set-theoretic concepts, thereby reducing the degree of misclassification of object data in images. In addition, a new bi-directional, self-organizing neural network (BDSONN) architecture, which incorporates fuzzy membership-based interconnections, is introduced. The enhanced performance of the BDSONN architecture, in respect of extraction times and extraction capabilities, is reported by means of a proposed system transfer index to reflect the noise immunity of the same. The BDSONN architecture is also able to segment multilevel image data if its transfer characteristics are guided by the multilevel MUSIG activation function. An efficient, parallel bidirectional self-organizing neural network (PBDSONN) architecture is also proposed for the segmentation and extraction of color images. From the application perspective, tracking of targets stands as an ideal example of soft computing-based object extraction. Methods based on neuro-fuzzy techniques have been used successfully to track both high- and low-speed moving objects from the motion scenes.

---

## Agent-based Management of Clinical Guidelines

**Candidate:** David Isern
**Institution:** Department of Computer Science and Mathematics, University Rovira i Virgili, Tarragona, Catalonia, Spain
**Supervisor:** Antonio Moreno
**Year awarded:** 2009
**URL:** http://www.tesisenxarxa.net/TDX-0313109-093946/

### Abstract

Clinical guidelines (CGs) contain a set of directions or principles to assist the healthcare practitioner with patient care decisions about appropriate diagnostic, therapeutic, or other clinical procedures for specific clinical circumstances. Guideline-based systems can constitute part of a knowledge-based decision support system in order to deliver the right knowledge to the right people in the right form at the right time. The automation of the guideline execution process is a basic step towards its widespread use in medical centres. This thesis focuses on the execution of CGs and proposes the implementation of an agent-based platform in which the actors involved in health care coordinate their activities to perform the complex task of guideline enactment. The internal elements of the platform act autonomously and proactively, and that improves the flexibility and robustness of the system, and allows coordination of the execution of tasks in an effective way. The implemented agent-based platform permits to represent all different roles and relationships (e.g. actors, medical devices, organization dependencies). A novel personalization method allows delivery of patient-oriented medical services. It maintains a user's profile that is employed to rate and rank alternatives composed by the platform, and an unsupervised learning method that allows the system to maintain dynamically this user's profile. Finally, the implementation of different application ontologies, which deal with all medical and organizational knowledge managed among all entities, permits the separation of the knowledge representation from its use (by agents). This approach allows description of declarative and procedural knowledge accurately. In addition, it allows adaptation of the system to different (execution) circumstances without changing the internal behaviour of agents. Moreover, this work studies the adequacy of intelligent agents in healthcare problems and the use of agent-oriented software engineering methodologies to implement (real) agent platforms.

---