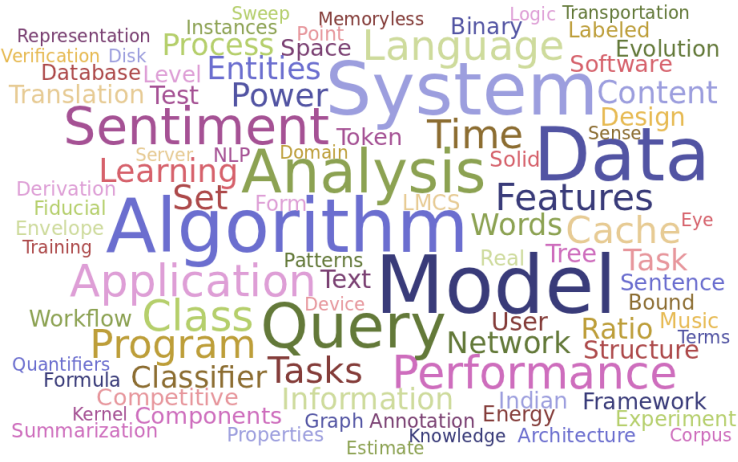
research@cse

A compilation of research abstracts



Department of Computer Science and Engineering
Indian Institute of Technology Bombay
February, 2014

Message from the Head

Welcome to the Research Scholar Poster Mela. The department of Computer Science and Engineering at IIT Bombay has had an impressive history leading computer science efforts in India, and a very good present, impacting computer science research internationally, and we believe it will have a great future as our research programs grow and improve. And the major factor in our future are our research scholars, who have grown in both numbers and in quality over the past years.

Today our department has internationally recognized groups in many areas. To name a few in alphabetical order: algorithms and complexity, data mining/machine learning, database systems, graphics, formal methods, information retrieval, natural language processing, and networks/distributed systems, as well as other smaller groups.

Our research work is today published in top tier (A/A+ level) conferences and journals, and many papers from IIT Bombay have been highly cited. Our faculty have received international recognition, as evidenced by the number of faculty who are/have been program chairs of top-tier conferences, and/or editors of top-tier journals.

Our research scholars have also won international recognition, winning fellowships and awards from leading companies such as Microsoft, Yahoo, IBM, TCS and Infosys, to name a few, with several of the awards coming from international competitions. The first ACM India doctoral dissertation award was won by Ruta Mehta last year.

But we must continue to aim higher. Our goal is to be recognized as not just the best or one of the best in India, but as one of the best overall in the world. With the efforts of our faculty and research scholars, supported by our masters students as well as bachelors students, I believe we can join the top 10 to 20 rank internationally in CSE within, say, a decades time.

So let me sign off, with kudos to our research scholars for their achievements, which we seek to showcase at the mela, and with best wishes for continued research success.

S. Sudarshan

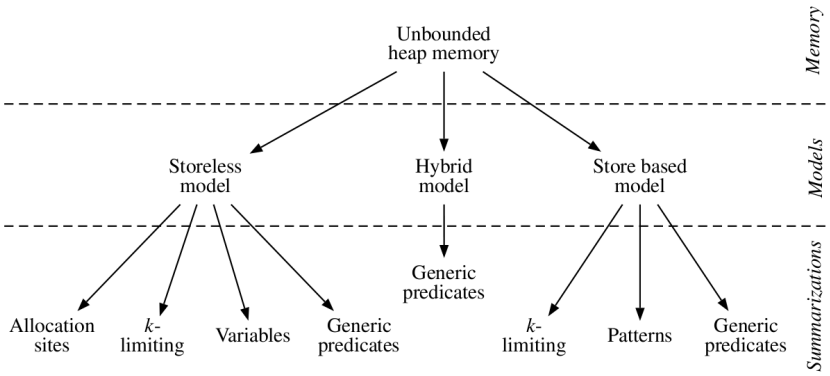
Contents

Heap Abstractions for Static Analysis

Vini Kanvar Prof. Uday Khedker

Study shows that the most common challenge faced by novice C/C++ programmers is the management of dynamic memory (heap). Understanding the concept of pointers in itself is nontrivial. Without this understanding, poorly written programs have memory leaks that impact the performance of the programs. Such programs use unnecessarily large system resources, and worst of all they fail due to out-of-resource problems. As a consequence, analysis of heap memory is becoming increasingly important.

Heap data is potentially unbounded and seemingly arbitrary. We provide a high-level view of the abstraction techniques used for static analyses of heap data. We view a *heap abstraction* as consisting of two features: a *heap model* to represent the heap memory and a *summarization* technique for bounding the heap representation. We classify the models as storeless, store-based, and hybrid. We review various summarization techniques, viz. k-limiting summarization, allocation-site based summarization, pattern based summarization, variable based summarization, and other generic predicates based summarization.



We have studied how these summarization techniques have been adopted in the literature for different heap models. As program analysts, we are still facing the challenge of creating summarizations that yield results that are both scalable to large sized programs and precise enough to produce useful results.

CONFIDE: Content-based Fixed-sized I/O Deduplication

Sujesha Sudevalayam Prof. Puru Kulkarni

Due to increased adoption of virtualization-based systems, there is a lot of inherent content similarity in systems like email servers, virtualized servers and file servers. Harnessing content similarity can help avoid duplicate disk I/O requests that fetch the same content repeatedly. This is referred to as I/O deduplication, the work in [Koller (2010)] maintains an additional content-based cache in order to serve more disk I/O requests from cache. It shows that any given fixed size cache can be better utilized if it is used as a content-based cache rather than a sector-based cache. However, by introducing an additional cache, it introduces cache exclusivity concerns among the caches. In our work, we incorporate intelligent I/O redirection within the storage virtualization engine of the device to manipulate the underlying sector-based cache itself as a content-based cache. We build and evaluate CONFIDE (Content-based Fixed-sized I/O Deduplication), a storage access optimization that identifies content similarity using fixed-sized blocks. The CONFIDE system builds upon the following key ideas: (i) VM images have a reasonably high level of similarity due to common golden master images [Peng (2012), Jayaram (2011)], (ii) Several application workloads like webmail proxy servers, mail servers and file servers have inherent content similarity [Koller (2010)], (iii) Using the entire cache as a content-based cache can increase overall cache efficiency. In this context, the contributions of our work are :-

1. Disk I/O de-duplication by content-deduplication over blocks.
2. Standard disk cache manipulated in a content-deduplicated fashion.

This work is applicable to any device containing a storage virtualization layer — SAN, volume controller, Shared object store, Hypervisor or VMM.

We build prototype implementation modules for both and perform trace-based evaluation in a simulation module. Results are promising, and throw light upon a couple of key points. First, while IODEDUP system reserves a part of the total memory to be used as a content-cache, the CONFIDE system manipulates the entire available buffer/disk cache space effectively like a content-cache. This is the key reason for better performance, up to 20% higher cache-hit ratio than IODEDUP for read/write traces. Second, although IODEDUP system uses Adaptive Cache Replacement (ARC) policy for its additional cache which boosts up its cache-hit ratio by 3-4 times [Koller (2010)] and hence performing better than CONFIDE for read-only traces, however CONFIDE system still performs better for read/write traces which are more representative of a real-world scenario. This also indicates that if the cache replacement policy could be changed from LRU to ARC in our system, then CONFIDE performance would also benefit from a similar performance boost. Although this may not be possible in a standard Linux host, ARC caches are already present in IBM's DS6000/DS8000 storage controllers [@ONLINE] and adoption of CONFIDE may help increase its cache performance.

Bibliography

[Koller (2010)] R. Koller and R. Rangaswami. I/O Deduplication: Utilizing Content Similarity to Improve I/O Performance. In *USENIX File and Storage Technologies (FAST)*, 2010.

[Peng (2012)] C. Peng, M. Kim, Z. Zhang, and H. Lei. VDN: Virtual Machine Image Distribution Network for Cloud Data Centers. In *INFOCOM, 2012 Proceedings IEEE*, pages 181 –189, March 2012.

[Jayaram (2011)] K. R. Jayaram, C. Peng, Z. Zhang, M. Kim, H. Chen, and H. Lei. An Empirical Analysis of Similarity in Virtual Machine Images. *Middleware 2011 Industry Track Workshop*.

[@ONLINE] http://en.wikipedia.org/wiki/Adaptive_replacement_cache

Holistic Optimization of Database Applications

Karthik Ramachandra Prof. S Sudarshan

Database backed applications today generate and operate on huge amounts of data. Most database applications are written using a mix of imperative language code and SQL. For example, database applications written in Java execute SQL queries through interfaces such as JDBC or Hibernate. Database stored procedures, written using languages such as PL/SQL and T-SQL, contain SQL queries embedded within imperative code. Further, SQL queries invoke user-defined functions (UDF s). UDF s can in turn make use of both imperative language constructs and SQL.

The imperative program logic is usually executed outside the database query processor. Queries are either embedded or dynamically constructed in the program, and are submitted (typically synchronously, and over the network), at runtime, to the database query processor. The query processor explores the space of alternative plans for a given query, chooses the plan with the least estimated cost and executes the plan. The query result is then sent back to the application layer for further processing.

A database application would typically have many such interactions with the database during its execution. Such interactions between the application and the database can lead to many performance issues that go unnoticed during development time. Traditional optimization techniques are either database centric (such as query optimization and caching), or application centric (such as optimizations performed by compilers), and do not optimize the interactions between the application and the database. This is because neither the programming language compiler nor the database query processor gets a global view of the application. The language compiler treats calls to the database as black-box function calls. It cannot explore alternative plans for executing a single query or a set of queries. Similarly, the database system has no knowledge of the context in which a query is being executed

and how the results are processed by the application logic, and has to execute queries as submitted by the application.

For example, repeated execution of parameterized queries and synchronous execution or queries result in a lot of latency at the application tier due to the many network round trips. At the database end, this results in a lot of random IO and redundant computations. Repeated execution of parameterized queries could be optimized to use set oriented query execution which saves both disk IO as well as network round trips. The effect of network and IO latency can be reduced by asynchronous submission of queries and asynchronous prefetching of query results. Performing such optimizations requires a thorough joint analysis of query and program. It cannot be conceived as purely query or purely program optimization technique.

We present holistic approaches [1, 2, 3, 4] which treat the database application as a unit, spanning the boundaries of the database and the application. Achieving this goal involves bringing together ideas from database query optimization, program analysis, and compilers, to analyze and transform the application. Our techniques are widely applicable to real world applications and can lead to significant improvement in performance.

Bibliography

[1] Karthik Ramachandra, Mahendra Chavan, Ravindra Guravannavar and S. Sudarshan: *Program Transformations for Asynchronous and Batched Query Submission*, CoRR abs/1402.5781, January 2014.

[2] Karthik Ramachandra, S. Sudarshan: *Holistic Optimization by Prefetching Query Results*, ACM SIGMOD 2012.

[3] Karthik Ramachandra, Ravindra Guravannavar and S. Sudarshan, *Program Analysis and Transformation for Holistic Optimization of Database Applications*, ACM SIGPLAN SOAP 2012 (*co located with PLDI 2012*).

[4] Ravindra Guravannavar and S Sudarshan: *Rewriting Procedures for Batched Bindings*, VLDB 2008.

Design of Blur Invariant Fiducial for Low Cost Quadcopter

Meghshyam Prasad Prof. Sharat Chandran Prof. Michael Brown

Introduction: A fiducial marker, or, simply a fiducial is an object placed in the field of view of an imaging system for use as a point of reference or a measure[1]. It is commonly used to track an object in an unknown environment and finds use in various applications in Virtual Reality, Medical imaging, Surveys, etc. ARToolkit [2], ARTag [3] are popular fiducials used in augmented reality applications.

Problem: The performance of these fiducials is satisfactory when there is little motion or no motion in the device obtaining the imagery. But when there is continuous and swift motion, as in the case of low cost quadcopters, performance of these fiducials degrade significantly due to motion blur.

Solution: Inspired from Circular Data Matrix[4] we have designed a fiducial that may be thought of as a binary code. It contains concentric white rings of equal widths on a black background with a blurred border. The outermost and innermost rings represent the start and end of the code and is embedded in the fiducial; these are not considered part of the code itself. The binary code is represented by the presence (or absence) of rings between "marker" rings. Depending on which ring is present or absent, the resulting binary code will change. The number of different patterns depends on the number of bits in the binary code. For example, if the binary code has three bits, there will be a maximum of three rings between"marker" rings and we end up with eight different patterns.

Algorithm: Our fiducial detection strategy is different from [4] and works under significant amount of blur. Our algorithm is based on the fact that there is no blur in the direction perpendicular to the direction of the motion. We apply a Gabor filter on the input image that detects the high gradient patches of the unblurred part of the fiducial. The output is clustered to

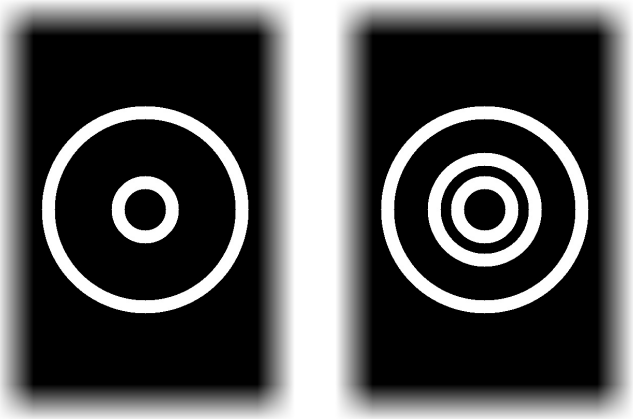


Figure 1: Sample two bit binary coded Fiducials (left: Binary Code 00, right: Binary Code 01)

group various patches pertaining to the fiducial. Principal Component Analysis (PCA) is now used to find the direction of grouped patches. We show that intensity profile along first principal component will give the "signature" of fiducial present in the image. Thus by comparing this "signature" with a standard, we are able to determine the binary code embedded in the fiducial.

Bibliography

[1] @ONLINE http://en.wikipedia.org/wiki/Fiducial_marker

[2] @ONLINE http://www.hitl.washington.edu/artoolkit

[3] Mark Fiala. ARTag, a Fiducial Marker System Using Digital Techniques, CVPR 2005.

[4] Leonid Naimark and Eric Foxlin. Circular Data Matrix Fiducial System and Robust Image Processing for a Wearable Vision-Inertial Self-Tracker, ISMAR 2002.

Learning to Collectively Link Entities

Ashish Kulkarni Kanika Agarwal Pararth Shah Prof. Ganesh Ramakrishnan

Search systems proposed today [1, 2] are greatly enriched by recognizing and exploiting entities embedded in unstructured pages. In a typical system architecture [3, 4] a spotter first identifies short token segments or "spots" as potential mentions of entities from its catalog. For our purposes, a catalog consists of a directed graph of typed nodes, to which entity nodes are attached. Many entities may qualify for a given text segment, *e.g.*, both *Kernel trick* and *Linux Kernel* might qualify for the text segment "...Kernel...". In the second stage, a disambiguator assigns zero or more entities to selected mentions, based on similarities between the contexts of the mention and the entity, as well as similarities between the entities. Collectively, these two stages comprise an *annotator*. An annotation record consists of a document ID, a token span, and one or more entities, e, chosen by the disambiguator. The success of an annotator depends on its ability to seamlessly bridge massive amounts of unstructured Web text with structured entity catalogs in a robust manner.

We present a system for robust disambiguation of entity mentions occurring in natural language text. Given an input text, the system aggressively spots mentions and their candidate entities. Candidate entities across all mentions are jointly modeled as binary nodes in a Markov Random Field. Their edges correspond to the joint signal between pairs of entities. This facilitates collective disambiguation of the mentions — achieved by performing MAP inference on the MRF in a binary label space. Our model also allows for a natural and graceful treatment of mentions that have either no corresponding entities or more than one correct entity. By restricting cliques to nodes and edges and with a submodularity assumption on their potentials, we get an inference problem that is equivalent to finding the min cut on a specially constructed flow network. We use a max margin framework, which is efficiently implemented using projected subgradient [5], for col-

lective learning. We leverage this in an online and interactive annotation system which incrementally trains the model as data gets curated progressively. We demonstrate the usefulness of our system by manually completing annotations for a subset of the Wikipedia collection. We have made this data publicly available. We also show that our node and edge features, along with our collective disambiguation approach, help achieve accuracy that is comparable to, or higher than existing systems.

Bibliography

[1] Chakrabarti S., Puniyani K., and Das S., *Optimizing scoring functions and indexes for proximity search in type-annotated corpora*, Proceedings of the 15th international conference on World Wide Web, ACM, New York, NY, USA, 2006.

[2] Kasneci G., Suchanek Fabian M., Ifrim G., Elbassuoni S., Ramanath M., and Weikum G., *NAGA: harvesting, searching and ranking knowledge*, Proceedings of the 2008 ACM SIGMOD international conference on Management of data, ACM, New York, NY, USA, 2008.

[3] Cucerzan, Silviu, *Large-scale named entity disambiguation based on Wikipedia data*, Proceedings of EMNLP-CoNLL, 708–716, 2007.

[4] Kulkarni S., Singh A., Ramakrishnan G., and Chakrabarti S., *Collective annotation of Wikipedia entities in web text*, Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '09, ACM Press, New York, New York, USA, 2009.

[5] Taskar, B., Chatalbashev, V., and Koller, D, *Learning associative Markov networks*, Proceedings of the twenty-first international conference on Machine learning, ACM, 2004.

Maximum Mean Discrepancy for Class Ratio Estimation: Convergence Bounds and Kernel Selection

Arun Iyer Prof. Sunita Sarawagi

Collaborator(s): Prof. Saketha Nath

The goal of this work [1] is to estimate the ratio of classes in any unlabeled test dataset given a labeled training dataset with an arbitrarily different ratio of classes. The closely related problem of creating a classifier with shifted class priors in the training and test set has been extensively studied. In contrast, our end goal is to estimate the ratio of classes and not the labels of individual instances in the unlabeled set. As an example consider websites that serve user directed content like news, videos, or reviews. Each item (article, video, or product) is associated with many user comments. An analyst wants to estimate the fraction of comments that express positive/negative sentiments. The polarity of each comment is not of interest.

A baseline estimator is to use the labeled data to train a classifier using supervised learning techniques and estimate class ratio of particular label y as the number of instances in the unlabeled data labeled with the label y. Since most supervised learning algorithms assume that the training and test data follow the same distribution, this method is unlikely to perform well. Many fixes for such a classifier have been proposed, however, the primary goal of these methods is to improve per-instance accuracy, and class ratios are estimated using the same paradigm of aggregating from per-instance predictions. Since we are not interested in the labels of individual instances, we explore direct methods of estimating the class ratios.

There have been three reported attempts for such direct estimation. The third, most recent approach [2], is based on minimizing the maximum mean discrepancy (MMD) over functions in a reproducing kernel Hilbert space (RKHS) induced by a kernel K. The MMD-based approach has several advantages: it is

applicable to arbitrary domains since it does not assume any parametric density on the data unlike conventional mixture modeling approaches. Because of this property, MMD has been successfully deployed in other problems including covariance shift and two-sample test. When deployed for class ratio estimation, it gives rise to a convex QP over a small number of variables, which is efficiently solvable. However, the approach has not been understood theoretically, empirical comparisons with other class ratio estimation methods are lacking, and kernel selection has not been adequately addressed.

In this work, we investigate the use of MMD-based approach for estimating such ratios. First, we theoretically analyze the MMD-based estimates. Our analysis establishes that, under some mild conditions, the estimate is statistically consistent. More importantly, it provides an upper bound on the error in the estimate in terms of intuitive geometric quantities like class separation and data spread. Next, we use the insights obtained from the theoretical analysis, to propose a novel convex formulation that automatically learns the kernel to be employed in the MMD-based estimation. Our kernel learning formulation turns out to be an instance of a Semi-Definite Program (SDP) with infinitely many linear constraints. To be able to work with infinite constraints, we design an efficient cutting plane algorithm for solving this formulation. We are aware of no prior work on kernel selection for this problem. We present an extensive evaluation of several existing methods, both from the direct and per-instance aggregation family, under varying true class ratios and training sizes. We obtain up to 60% reduction in class ratio estimation errors over the best existing method.

Bibliography

[1] Iyer, A, Saketha Nath, J, Sarawagi, S, Maximum Mean Discrepancy for Class Ratio Estimation: Convergence Bounds and Kernel Selection, ICML 2014.

[2] Zhang, K, Schölkopf, B, Muandet, K and Wang, Z, Domain adaptation under Target and Conditional Shift, ICML 2013.

Active Evaluation of Classifiers on Large Datasets

Arun Iyer Prof. Sunita Sarawagi

Collaborator(s): Namit Katariya

In this paper [1] we address the problem of evaluating the accuracy of a classifier $C(\mathbf{x})$ when deployed on a very large unlabeled dataset D. We are given a small labeled test set L. We consider situations where D is so large in comparison to L that the average accuracy over L is unlikely to be a reliable estimate of the classifier's real performance on the deployment data D. We present a method for more reliable accuracy estimates of $C(\mathbf{x})$ on D using the given labeled set L, and a method for selecting additional instances to be labeled to further refine the estimate.

This problem has applications in many modern systems that rely on the output of imperfect classifiers. Consider a concrete example. A search engine needs to deploy a classifier $C(\mathbf{x})$ to label if a Web page $\mathbf{x}$ is a homepage. Since the label is used to decide on the rank of a web page in a search result, it is important to calibrate reliably the accuracy of the classifier on a general web corpus D. Typically, editors hand pick a set of instances L, label them as homepage or not, and measure the accuracy of $C(\mathbf{x})$ on L. This method is likely to be flawed because the Web is so diverse and huge that it is difficult for humans to select a representative set while keeping L small.

In spite of the practical importance of the problem, existing work on the topic is surprisingly limited. The standard practice in classifier evaluation is to use a fixed labeled test set to evaluate a classifier which is then deployed for predictions on 'future' instances. Can we improve this process when the deployment set is available in unlabeled form? The use of unlabeled data for *learning* a classifier has received a lot of attention in the context of topics like active learning, semi-supervised, and transductive learning. However, the task of *learning* a classifier is very different from the task of *evaluating* a given classifier. The only

existing work on selecting instances for evaluating a classifier are: [3] which presents a new proposal distribution for sampling, and [2, 4] which use stratified sampling. Both these methods assume that the classifier $C(\mathbf{x})$ is probabilistic and their selection is based solely on the classifier's $\Pr(y|\mathbf{x})$ scores. We wish to evaluate the accuracy of any classifier: be it a set of manually developed rules, a learned generative model, or a non-probabilistic method like a decision tree — none of these can be handled by the methods in [3, 2, 4].

Our method is founded on the principles of stratified sampling like in [2, 4] but with important differences. Instead of *fixing* a stratification, we *learn* a stratification in terms of a generic feature space and the strategy evolves as more data gets labeled. For stratifying data, we use hashing instead of conventional clustering based approaches as the latter do not scale well. We design a novel algorithm for learning hash functions that cluster instances with similar accuracies more effectively than existing learning techniques for distance preserving hashing. Also, none of the existing estimation methods consider the case where the dataset D is so large that even a single sequential scan over it will take hours. Our method is designed to perform accuracy estimation and instance selection on D which can only be accessed via an index. We achieve upto 62% reduction in relative error over existing methods in our experiments with datasets that range in sizes from 0.3 million to 50 million instances.

Bibliography

[1] Katariya, N, Iyer, A, Sarawagi, S, Active Evaluation of Classifiers on Large Datasets, ICDM 2012.

[2] Bennett, P, Carvalho, V, Online stratified sampling: evaluating classifiers at web-scale, CIKM 2010.

[3] Sawade, C, Landwehr, N, Bickel, S, Scheffer, T, Active Risk Estimation, ICML 2010.

[4] Druck, G, McCallum, A, Toward interactive training and evaluation, CIKM 2011.

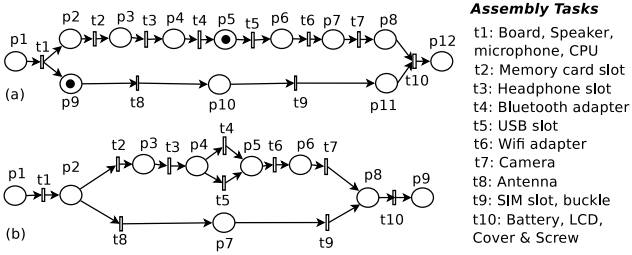
Token Transportation in Petri Net Models of Workflow Patterns

Ahana Pradhan Prof. Rushikesh K. Joshi

Real world business processes often need to evolve. The separation between *build time* and *run time* in the traditional workflow system architecture makes dynamic workflow evolution a challenge. The reason is primarily that it demands evolution capabilities in both the phases. A change to a *build time* artifact (process specification) needs to be applied to all existing running instances of the process, which may all be in different runtime states. As a consequence, a workflow evolution assignment needs to address questions such as whether a given process should be continued as it is or restarted, or can it be migrated to a new business logic without the need to relaunch.

When a business logic needs to be migrated into a new logic, it may be possible to reuse some of the existing completed tasks and yet let the process resume under the changed process flow. However, the exact placement of resumption in the new flow depends on the current placements of the tokens in the Petri net model of the workflow, the logic that follows, and the logic that precedes the tokens. This problem of smoothly migrating tokens from an old workflow into a new workflow is addressed. In particular, we consider various cases of token migration, which are formulated in terms of token migration among a set of workflow patterns described in [2]. These formulations are referred to as primitive token transportation patterns. We show that the patterns can be further used to carry out token transportation in larger nets through a strategy called *yo-yo* transportation. We introduce the notion of *token transporter bridges*. A token transportation bridge is a modular unit for token transportation, and it captures token transportation for all markings of a given WF-net. The modular bridges are composed of inter-related flow-jumpers, which were introduced by Ellis et. al [1] . The work also deals with the issues of structural constraints and the consistency criteria for migration.

An example of a workflow migration case is given in the figure below. Fig. (a) depicts a non-primitive product-line workflow process of a single product that is to be evolved into a workflow process in Fig. (b) delivering three separate products. The old workflow assembles a heavy featured tablet phone device, which is to be morphed into three low cost separate products, a simple phone, a tablet with a USB and a tablet with Blue-tooth. The token migration strategy corresponds to migration of live incomplete assemblies into one of the three products whenever possible without removing the parts which have been already assembled. The last constraint corresponds to consistency criteria. The marking in the old net shows completion of tasks t_1, t_2, t_3 and t_4. In the poster the Yo-Yo approach to token transportation and the concept of token transportation bridges are described with the help of this example.



Migration Case: (a) Old Net (b) New Net

Bibliography

[1] C. A. Ellis and K. Keddara. A workflow change is a workflow. In Business Process Management, Models, Techniques, and Empirical Studies, pages 201-217. Springer-Verlag, 2000

[2] W. M. P. Van Der Aalst, A. H. M. Ter Hofstede, B. Kiepuszewski, and A. P. Barros. Workflow patterns. Distrib. Parallel Databases, 14(1):5-51, jul 2003

Modeling Component Interactions and Evolution in Software Architecture

Dharmendra K. Yadav Prof. Rushikesh K. Joshi

Most software systems are subjected to changes due to continuously changing user requirements. To incorporate these changes, the artifacts in the software must evolve. One of the key artifacts in software is its architecture. It describes the software at the highest level of abstraction. Evolving software architecture to accommodate new features, at the same time ensuring preservation of earlier properties becomes a complex task. To perform the task of evolution rigorously, a method based on modeling component interactions is proposed. In this method, software architecture (component and connector view[1]) has been modeled using CCS (Calculus of Communicating Systems)[2]. In these models, components are primary entities having identities in the system and connectors provide the means for enablers of interactions i.e. behaviours between them. This view is very similar to the abstractions provided by CCS, in which, components can be seen as non-movable agents and connectors as complex interactions. The CCS effectively captures interactions occurring at architectural level through its features such as components and their compositions, input/output actions over channels and non-deterministic behavior. The use of CCS in software architectural modeling has earlier been demonstrated in [3]. Desired properties in the model are ensured through property verification.

The method proposes evolution of software architecture through reuse-centric model transformation. The transformation involves making minor variations and reconnections among original components along with introduction of additional components. An illustration of a technique based on restriction and introduction of new agent is presented in Figure 2. In the example, a system is evolved through relabeling the ports of components, achieving decoupling with other components with which they interact. Further, as new requirements may demand introduction of new components for evolution of the earlier model of architecture, new

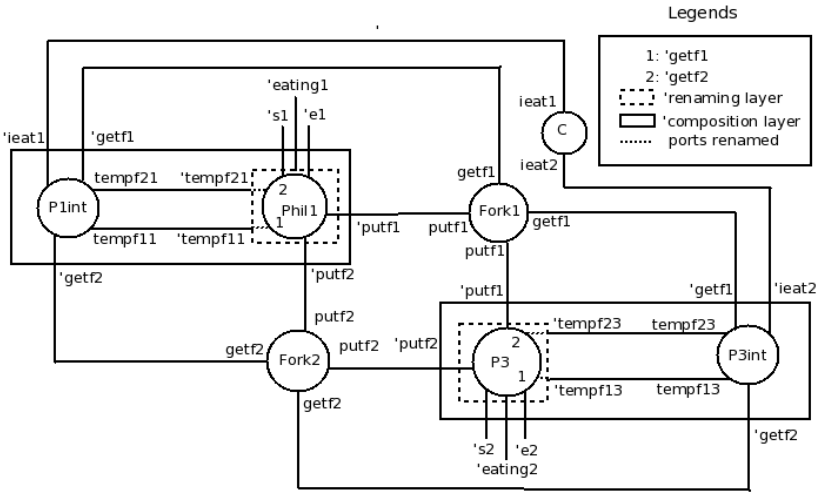


Figure 2: **The Evolution Process**

components having ports matching with the ports of existing components of the architecture can also be added. Restriction makes actions unavailable for further composition. The proposed method models software architecture and its evolution within the framework of CCS. The focus is on developing a methodology of modeling software architectural components and interactions in CCS and carrying out reuse-centric evolution of these models.

Bibliography

[1] Perry, D.E., Wolf, A.L.: Foundations for the study of software architecture. ACM SIGSOFT Software Engineering Notes **17**(4) (1992) 40–52

[2] Milner, R.: Communication and Concurrency. Prentice-Hall (1989)

[3] Yadav, D., Joshi, R.: Capturing interactions in architectural patterns. In: Proceedings of the 2nd IEEE conference on Advance Computing Conference (IACC),. (2010) 443-448

Practical Quantifier Elimination for Linear Bit-Vector Arithmetic

Ajith K John Prof. Supratik Chakraborty

Existential quantifier elimination (QE) is the process of transforming a formula containing existential quantifiers into a semantically equivalent quantifier-free formula. For example, the formula $\exists x.\,((x - y \leq 0) \wedge (y + 2z \leq x))$ is equivalent to the quantifier-free formula $(2z \leq 0)$ in the theory of integers. This has a number of important applications in formal verification and program analysis. Verification and analysis tools often assume unbounded data types like `integer` or `real` for program variables, and use QE techniques for unbounded data types, for ease of analysis. However, the results of analysis assuming unbounded data types may not be sound if the implementation uses fixed-width words, and if the results of operations overflow without being detected. This motivates us to investigate QE techniques for constraints involving fixed-width words (bit-vectors). Specifically, we are interested in techniques for QE from Boolean combinations of bit-vector (linear modular) equalities, disequalities and inequalities.

Conversion to bit-level constraints (bit-blasting) [1] followed by bit-level QE is arguably the dominant technique used in practice for eliminating quantifiers from linear modular constraints (LMCs). This approach, though simple, destroys the word-level structure of the problem and does not scale well for LMCs with large modulus. Since LMCs can be expressed as formulae in Presburger Arithmetic (PA) [1], QE techniques for PA (e.g. those in [2]) can also be used to eliminate quantifiers from LMCs. However, converting the results obtained as PA formulae back to LMCs is often difficult. Also, empirical studies have shown that using PA techniques to eliminate quantifiers from LMCs often blows up in practice [1].

We present a practically efficient QE algorithm for conjunctions of LMCs [3, 4]. We use a layered approach, i.e., sound but

relatively less complete, cheaper layers are invoked first, and expensive but more complete layers are called only when required. The first layer involves elimination of quantifiers from the given conjunction of LMCs and simplification of the conjunction using the equalities present in conjunction. The second layer employs an efficient heuristic to drop unconstraining inequalities and disequalities in the conjunction. The third layer is an extension of Fourier-Motzkin QE algorithm for LMCs, and the final (expensive) layer involves model enumeration. We present approaches to extend this algorithm to work with Boolean combinations of LMCs. Experiments demonstrate the effectiveness of our algorithm over alternative QE techniques based on bit-blasting and PA. In all our experiments, most quantifiers were eliminated by the cheaper layers, which underlines the importance of our layered framework for QE.

Bibliography

[1] D. Kroening, O. Strichman. *Decision procedures : an algorithmic point of view*, Texts In Theoretical Computer Science, Springer 2008

[2] W. Pugh. *The Omega Test: A fast and practical integer programming algorithm for dependence analysis.* Communications of the ACM, Pages 102-114, 1992

[3] A. John, S. Chakraborty. *A quantifier elimination algorithm for linear modular equations and disequations*, In CAV 2011

[4] A. John, S. Chakraborty. *Extending quantifier elimination to linear inequalities on bit-vectors*, In TACAS 2013

Formal methods for analysis of biological systems

Sukanya Basu Prof. Supratik Chakraborty Prof. Akshay S Prof. Ashutosh Trivedi

NFkB is a key transcription factor involved in deciding the fate of a cell –pro-survival or apoptotic (pro-death) [1]. Some recent experiments conducted by our collaborators at ACTREC, Tata Memorial Centre, suggest that PSMD9 and PSMD10 proteasomal subunits are engaged in a potential switch-like control to regulate NFkB activity. It is essential to investigate PSMD9 and PSMD10 protein-interaction networks, and discern their control of NFkB-mediated cell fate decisions.

Performing extensive biological experiments are highly time and resource consuming. We plan to integrate data from well-curated structural and functional networks and from results available in the literature. The real crux of the objective, which is the cross talk between pro-survival and apoptotic pathways, can only be addressed by creating a network using molecules known to be involved in these pathways.

There are two key challenges from a computational point of view. First, the level of curation is likely to vary significantly in different parts of the network. Hence, techniques that allow us to reason in the presence of varying confidence measures need to be developed and fine-tuned. Second, in any given part of the network, there is inherent imprecision and noise in the quantitative annotation that one can obtain from experimental data. This warrants the use of modeling and analysis techniques that are robust under imprecise and noisy data.

We propose to incorporate data from various publicly available databases documenting interactions and regulatory pathways [2, 3, 4] to create a “master”network. We would then adopt an approach that starts with the entire “master”network and zooms

down on the relevant sub-networks guided by confidence and precision metrics, and objectives of different network queries. Analyzing the sensitivity of the produced results on the confidence and precision of different sub-networks would allow us to calibrate the parameters of the model better, and also provide feedback for designing targeted biochemical and genetic experiments. The resulting networks will be interrogated using microarray expression profiles obtained from PSMD9 and PSMD10 overexpression with the following specific queries:

1. How are the apoptotic or prosurvival pathways influenced by the genes that are up or down regulated by overexpression of PSMD9 and PSMD10?

2. Does the network provide a probable understanding of the opposing effects of PSMD9 and PSMD10 on NFkB activity? If so, does that lead to quantifiable differences?

3. How do we design targeted genetic experiments, and obtain quantitative data which will improve confidence in the model and in the answers to the above questions?

Bibliography

[1] Vijay R. Baichwal and Patrick A. Baeuerle; Apoptosis: Activate NF-kB or die? Current Biology 1997, 7:R94-R96

[2] Kanehisa, M. and Goto, S.; KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Res. 28, 27-30 (2000).

[3] Arntzen MØ, Thiede B. ApoptoProteomics, an integrated database for analysis of proteomics data obtained from apoptotic cells. Mol Cell Proteomics. 2012

[4] Vastrik I, D'Eustachio P, Schmidt E, Joshi-Tope G, Gopinath G, Croft D, deBono B, Gillespie M, Jassal B, Lewis S, Matthews L, Wu G, Birney E, Stein L. Reactome: a knowledge base of biologic pathways and processes. Genome Biology 2007.

Smart-Phone Based Effective Teaching in Large Classroom Settings

Mukulika Maity Prof. Bhaskaran Raman Prof. Mythili Vutukuru

Worldwide, the demand for quality education is high, unfortunately far exceeding its supply. While an ideal teacher:student ratio is around 1:20 to 1:30, in developing countries like India class sizes of around 100 students are common-place, and often touching 500 to 1,000 students! On the other hand, the smartphone/tablet revolution is fast making inroads into the student population. So, our goal is to enable effective teaching using mobile platforms in large classroom settings.

This entails network and system level challenges. The networks challenges are -(1) First, the density of networked smartphone usage is unique and high (2) The traffic pattern is unique too. In many applications, all users will be using the network simultaneously and similarly. (3) In applications like lecture quiz, the completion time of the *last user* is an appropriate metric, not the median completion time. The system issue is the instructor must have control over students devices in classroom i.e. maintaining student integrity. Often times it is required to enforce that users are using *only* a particular application, and no other. Without this, smart-devices in a classroom can be counter-productive.

To study the effect of the WiFi in dense settings, we performed detailed measurement studies in controlled settings as well as in a real classroom setting, with the help of students taking a course over several weeks. We experienced poor performance in classroom as the aggregate throughput was very low compared to prior studies[1],[2]. In case of lecture quiz, the worst case completion time was really high. We also reproduced a similar setting in our laboratory with volunteers, albeit in a more controlled fashion, to better understand some of the network effects we saw in the live class-room. We observed that increased contention due to HTTP-chattiness causes poor performance of in-class applications, an aspect not considered in prior dense-WiFi studies. We

have come up with a simple two-class flow scheduler that restricts the number of active flows at any instant of time to reduce channel contention.

We have designed an authentication and control framework to act as a deterrent to violating integrity. The framework (a) monitors the client device (monitors activity on smart device, disables all notifications and monitors connectivity to 3G, Bluetooth etc.) to provide information to the instructor in the class- room, and (b) includes a light-weight camera-based authentication mechanism to help determine that the device is actually being used by the student that claims to be using it. We have prototyped the authentication framework on the Android platform, and performed several experiments with students emulating cheating scenarios.

The demand for quality education is much higher today than its supply. We believe bringing together several mature and developing mobile and wireless technologies has enormous potential for improving the quality of education.

Bibliography

[1] S. Choi, K. Park, and C. kwon Kim. On the Performance Characteristics of WLANs: Revisited. In SIGMETRICS, Jun 2005.

[2] M. A. Ergin, K. Ramachandran, and M. Gruteser. An experimental study of inter-cell interference effects on system performance in unplanned wireless LAN deployments. Computer Networks, 52, 2008.

A Functional Approach for Flow and Context Sensitive Pointer Analysis

Pritam Gharat Prof. Uday Khedker

Static program analysis consists of automatically discovering the properties of a program at compile time by examining the static representation of that program. These properties can be used for program verification, validation, testing, optimizations.

The analysis gets complicated for languages that support pointers. In order to analyze a program that involves pointers, it is necessary to know what each variable may point to at a given program point. It has been proved that a single procedural points-to analysis with dynamic memory is undecidable, even when only scalar variables are allowed.

The analysis gets further more complicated when an application consists of multiple procedures. This brings in the need for interprocedural analysis. An interprocedural data flow analysis incorporates the effect of callers on callees and vice-versa.

Safety of an analysis can be ensured by covering all valid execution paths; excluding a valid path may result in an unsafe solution. However, including an invalid path may lead to imprecise results. Hence, a precise interpocedural analysis considers only interprocedurally valid paths which have proper call-return matching.

There has been a belief that precision and efficiency cannot go hand in hand. Several approaches try to achieve efficiency at the cost of precision. However, precision and efficiency both can be achieved by computing only the relevant information. We aim to achieve efficiency by analyzing a procedure only once and compute only the information that will be used in the future by making it liveness driven.

Grammatical Error Correction with Applications to Post-Editing of MT output

Anoop Kunchukuttan Prof. Pushpak Bhattacharyya

Collaborator(s): Ritesh Shah

We investigate methods for correction of grammatical errors in text, especially those written by non-native speakers of English. Non-native speakers make typical errors like determiner, noun number, subject-verb agreement, verb form and preposition errors. The problem is particulary challenging because the data is skewed (typically $< 5\%$ of the text contains grammar errors). We present our grammar correction system which corrects noun-number, determiner and subject-verb agreement errors. For noun-number and determiner correction, we apply a classification approach using rich lexical and syntactic features. For subject-verb agreement correction, we propose a new rule-based system which utilizes dependency parse information and a set of conditional rules to ensure agreement of the verb group with its subject. Our system obtained an F-score of 21.41% on the CoNLL-2013 shared task test set using the M^2 evaluation method.

A sentence may contain multiple interacting grammatical errors. Most methods for joint correction learn local correction models, but perform a joint inference. In contrast, we explore joint modelling of some dependent grammatical error types.

Such grammatical errors are common in machine translation output. We are exploring methods to correct these grammatical errors by learning corrections from monolingual target language corpus. We will investigate how source sentence information can help disambiguate among multiple possible corrections, a feature that distinguishes this work from grammatical error correction in non-native text.

Bibliography

[1] Anoop Kunchukuttan, Ritesh Shah, Pushpak Bhattacharyya. IITB System for CoNLL 2013 Shared Task: A Hybrid Approach to Grammatical Error Correction. CoNLL 2013 .

Generalizations of the Łoś-Tarski Preservation Theorem

Abhisekh Sankaran Prof. Bharat Adsul Prof. Supratik Chakraborty

We present new preservation theorems that semantically characterize the $\exists^k\forall^*$ and $\forall^k\exists^*$ prefix classes of First Order Logic (FOL) formulae, for each natural number k. Unlike preservation theorems in the literature that characterize the $\exists^*\forall^*$ and $\forall^*\exists^*$ prefix classes as a whole, our theorems provide finer characterizations by relating the count of quantifiers in the leading block of the quantifier prefix to natural quantitative properties of models. As special cases of our results, we obtain the classical Łoś-Tarski preservation theorem for formulae, in both its forms, substructural and extensional. We extend these results to provide, for each natural number n, semantic characterizations of the subclasses of the Σ^0_n and Π^0_n prefix classes of FOL formulae, in which the number of quantifiers in the leading block of the quantifier prefix is fixed to a given natural number k. These extensions are new preservation theorems that give finer (in a similar sense as mentioned earlier) characterizations of the Σ^0_n and Π^0_n prefix classes than those in the literature.

We also give a new semantics to the literature notion of 'relativization' and show that using this semantics, the Łoś-Tarski preservation theorem, for relational vocabularies, can be stated in an alternate form that yields more insight into the structure of a characterizing universal, resp. existential, sentence for a sentence that is preserved under substructures, resp. extensions. Our notion of relativization also serves as a meta-technique for showing that our preservation theorems hold over interesting classes of finite structures like words, trees, co-graphs and acyclic graphs of bounded diameter.

Bibliography

[1] A. Sankaran, B. Adsul and S. Chakraborty, *A Generalization of the Łoś-Tarski Preservation Theorem over Classes of*

Finite Structures, http://arxiv.org/abs/1401.5953, January 2014.

[2] A. Sankaran, B. Adsul and S. Chakraborty, *Generalizations of the Łoś-Tarski Preservation Theorem*, http://arxiv.org/abs/1302.4350, June 2013.

[3] A. Sankaran, B. Adsul, V. Madan, P. Kamath and S. Chakraborty, *Preservation under Substructures modulo Bounded Cores*, WoLLIC 2012, Springer, pp. 291-305.

Improving the Energy Efficiency of MapReduce Framework

Nidhi Tiwari Prof. Umesh Bellur Prof. Maria Indrawan
Dr. Santonu Sarkar

Collaborator(s): IITB-Monash Research Academy, Infosys Ltd.

The increase in size of the data-centers to support a high number of users, has led to many-fold increase in their energy consumption. Many of these data-centers contain Hadoop MapReduce clusters of hundreds and thousands of machines, to process the infrequent batch and interactive big data jobs. Such an organization makes the MapReduce clusters energy inefficient as a large number of machines are underutilized for long time. Our project aims to improve the energy efficiency of MapReduce clusters. To realize this we will do a comprehensive energy characterization of Hadoop MapReduce and create its energy-performance model. The what-if analysis done using this model will help in configuring an energy efficient Hadoop MapReduce. The energy characterization will be used to derive heuristics for designing energy aware scheduling algorithm. The energy aware configuration and scheduling will improve the energy efficiency of MapReduce clusters thus help in reducing the operational costs of the data-centers.

Using Exchange Argument to Prove Query Lower Bounds

Jagadish M. Prof. Sridhar Iyer

'Exchange Argument' is one of the techniques used to prove the correctness of greedy algorithms. The technique is usually taught at any introductory course on algorithms. Although many undergrad textbooks discuss this technique, only a handful of illustrative examples like Minimum Spanning Tree and Interval Scheduling are found across different texts ([2],[1], [3]). We show that the exchange argument can also illustrated in a different context of proving query lower bounds.

We briefly describe the query model of computation and provide context for our work. The query-model or decision-tree model is a computational model in which the algorithm has to solve a given problem by making a sequence of queries which have 'Yes' or 'No' answers. A large class of algorithms can be described on this model and we can also prove non-trivial lower bounds for many problems on this model. We refer to lower bounds on this model as 'query lower bounds'.

Many lower bounds on the query-model are proved using a technique called adversary argument. In CS courses, a common example used to illustrate the adversary argument is the following problem: Suppose there is an unweighted graph G with n vertices represented by an adjacency matrix. We want to test if the graph is connected. How many entries in the adjacency matrix do we have to probe in order to test if the graph has this property (property being 'connectivity')? Each probe is considered as a query.

Since the adjacency matrix has only n^2 entries, $O(n^2)$ queries are sufficient. It is also known that $\Omega(n^2)$ queries are necessary. Proving this lower bound is more difficult and is done using the adversary argument [3].

In literature, we find that lower bound proofs of this problem rely too much on 'connectivity' property and do not generalize well ([3],[4]). When the property being tested is changed, the proof changes significantly. We show that the exchange argument

gives a more systematic way of proving lower bounds for problems involving testing of graph-properties. We did a pilot experiment and found that students were able to understand and apply our method.

Bibliography

[1] Cormen, Thomas H., et al. Introduction to Algorithms. Vol. 2. Cambridge: MIT press, 2001.

[2] Kleinberg, Jon, and Éva Tardos. Algorithm Design. Pearson Education India, 2006.

[3] Jeff Erickson. Algorithms. (Lecture Notes) http://www.cs.uiuc.edu/∼jeffe/teaching/algorithms/ (Last Accessed: 14 Feb 2014).

[4] Arora, Sanjeev, and Boaz Barak. Computational Complexity: A Modern Approach. Cambridge University Press, 2009.

Precise Call Disambiguation in the Presence of Function Pointers

Swati Rathi Prof. Uday Khedker

Interprocedural data flow analysis increases the precision of data flow information by incorporating the effect of callers on callees and vice-versa. Thus it requires as input, the caller-callee relationships. In presence of calls through function pointers, discovering exact caller-callee relationship may not be possible at compile time and hence they have to be approximated. Safety of interprocedural analysis requires that caller-callee relationships should not be under-approximated, its precision requires that these relationships should not be over-approximated. Also if we over-approximate the callee information, we will have to consider additional callees. This will not only compute imprecise results but also affects the efficiency of the analysis. Hence, precise call disambiguation is essential for precise and efficient interprocedural analysis.

Learning a Bounded-Degree Tree Using Separator Queries

Jagadish M. Anindya Sen

In the context of learning from data the structure of an undirected graphical model, there exists a class of approaches called *constraint-based structure learning methods*. These approaches attempt to construct a network structure that best captures the dependencies in the data. Learning happens via *conditional independence* (**CI**) queries (also called separator queries) and various statistical tests are employed to answer the queries. For our present work, we model the statistical-test-based **CI** query as follows. Suppose we are given a hidden graph $G = (V, E)$, where the vertex set V corresponds to the set of variables in the dataset. We assume the existence of a *perfect* oracle, which when asked queries of the form $(X \perp\!\!\!\perp Y | Z)$?, where $X, Y, Z \subset V(G)$ are three disjoint sets of vertices, returns 'Yes' if removing all vertices in Z disconnects subgraphs $G(X)$ and $G(Y)$; and 'No' otherwise. In other words, Z *separates* X from Y. The task is to infer $E(G)$ from answers to separator queries.

For general graphs, the problem is NP-hard and its hardness is proportional to the tree-width of the graph. Tree-width of a graph is a measure of how close its structure is to that of a tree. Graphs with high tree-width contain large cliques and are therefore harder to learn. Hence, it is useful to focus on graphs with low tree-width. The current work focuses on the simplest of such structures *i.e.* graphs with tree-width one (or trees). For trees, the general **CI** query can be simplified by restricting X, Y and Z to singleton sets. This is because an arbitrary **CI** query on trees can be broken down into a number of less expensive "simple" queries of the form: "Does the node y lie on the path between node x and node z?". Answering each restricted query now takes constant time. The objective is to minimize the time taken to output the tree in terms of n.

If the tree has unbounded degree, the problem can be solved in $\Theta(n^2)$ time. Our main result shows that it is possible to im-

prove the upper bound for bounded-degree trees. To the best of our knowledge, no $o(n^2)$ algorithm is known even for constant-degree trees. We also give an $O(d^2 n \log^2 n)$ randomized algorithm and prove an $\Omega(dn)$ lower bound for the same problem. We leave the problem of obtaining an $O(n^{1.5-\epsilon})$ tree-structure learning algorithm, if one exists, open.

Bibliography

[1] Jagadish, M., Sen, A., Learning a Bounded-Degree Tree Using Separator Queries. In: Proceedings of the 24th International Conference on Algorithmic Learning Theory (ALT), pp. 188-202, 2013.

[2] D. Koller and N. Friedman, Probabilistic Graphical Models: Principles and Techniques. MIT Press, 2009.

Query Interpretation and Response Ranking for Entity-aware Search

Uma Sawant Prof. Soumen Chakrabarti

Recent studies show that a large fraction of Web search queries revolve around *entities*. Entities are real life objects of interest: people, locations, electronic goods and many more. However, large part of the Web is still unstructured lacking any indication of its connection with entities. Large scale linking of entity mentions in the corpus with corresponding entities will allow us a richer view of the data. For example, a sentence "Master blaster answers critics with his bat, slams century in Ranji match" can be interpreted as Master blaster[Sachin Tendulkar, person] answers critics with his bat, slams century in Ranji[Ranji Trophy, cricket competition] match". Once entities are identied, one can additionally think in terms of types of entities and relations between entities.

In this work[2, 1], we propose to design algorithms which can possibly create and utilize such entity-linked data and respond

to user queries in terms of documents or entities (as answers to user queries) or as related entities (not as answers, but to facilitate browsing and to increase user engagement). We refer to this kind of search as *Entity-aware search*. Entity and entity type annotation for corpus and some very popular queries can be done a priori as they are available offline. However, many of the user queries need to be interpreted dynamically using the context such as words in the query, other queries in the query session, click data, user demographics and so on. Given a particular interpretation of user query, the search algorithm needs to perform efficient query-dependent aggregation of information available from across the corpus thanks to a priori entity linking. Additionally, the search algorithm is required to be robust to annotation noise, as large-scale entity annotations are machine learnt and not 100% accurate. Our work lies in exploring all these aspects and its effect on response ranking.

Prior research follows a two-staged process for query answering: interpretation of user queries followed by response ranking over the corpus or a structured knowledge base. In contrast, we incorporate the signal from the corpus into query interpretation, effectively making it a *joint query interpretation and ranking approach*. For a class of queries which are seek entities by specifying a target type (e.g. losing team baseball world series 1998), we show that such a joint approach provides superior performance over two-staged approach. We present two different models for realizing this joint framework, one based on generative language models and the other on a discriminative max-margin framework.

Bibliography

[1] Uma Sawant, Soumen Chakrabarti, Learning Joint Query Interpretation and Response Ranking. In proceedings of WWW 2013, Rio De Janeiro, Brazil.

[2] Uma Sawant, Soumen Chakrabarti, Features and Aggregators for Web-scale Entity Search. Technical report, 2013.

Structure Cognizant Pseudo Relevance Feedback

Arjun Atreya V Prof. Pushpak Bhattacharyya Prof. Ganesh Ramakrishnan

Collaborator(s): Yogesh Kakde

We propose a structure cognizant framework for pseudo relevance feedback (PRF). This has an application, for example, in selecting expansion terms for general search from subsets such as Wikipedia, wherein documents typically have a minimally fixed set of fields, *viz.*, *Title, Body, Infobox* and *Categories*. In existing approaches to PRF based expansion, weights of expansion terms do not depend on their field(s) of origin. This, we feel, is a weakness of current PRF approaches.

We propose a per field EM formulation for finding the *importance* of the expansion terms, in line with traditional PRF [1]. However, the final weight of an expansion term is found by weighting these *importance* based on whether the term belongs to the title, the body, the infobox or the category field(s). In our experiments with four languages, *viz.*, English, Spanish, Finnish and Hindi, we find that this structure-aware PRF yields a 2% to 30% improvement in performance (MAP) over the vanilla PRF.

We conduct ablation tests to evaluate the importance of various fields. As expected, results from these tests emphasize the importance of fields in the order of title, body, categories and infobox.

Bibliography

[1] Chengxiang Zhai and John Lafferty. 2001. Model-based feedback in the language modeling approach to information retrieval. In *Proceedings of the tenth international conference on Information and knowledge management*, CIKM '01, pages 403–410, New York, NY, USA. ACM.

A Computational Framework for Boundary Representation of Solid Sweeps

Prof. Bharat Adsul Jinesh Machchhar Prof. Milind Sohoni

This paper is about the theory and implementation of the solid sweep as a primitive solid modeling operation. A special case of this, viz., blends is already an important operation and prospective uses for the sweep are in NC-machining verification [1, 3, 8, 9], collision detection, assembly planning [1] and in packaging [10].

The solid sweep is the envelope surface $\mathcal{E}$ of the swept volume $\mathcal{V}$ generated by a given solid M moving along a one-parameter family h of rigid motions in $\mathbb{R}^3$. We use the industry standard brep format to input the solid M and to output the envelope $\mathcal{E}$. The brep of course has the topological data of vertices, edges and co-edges, loops bounding the faces and orientation of these, and the underlying geometric data of the surfaces and curves. As we show, the brep of $\mathcal{E}$, while intimately connected to that of M, has several intricate issues of orientation and parametrization.

Much of the mathematics of self-intersection, of passing body-check and of overall geometry have been described in the companion paper [6] . This paper uncovers the topological aspects of the solid sweep and its construction as a solid model. Here, we restrict ourselves to the simple generic case, i.e., smooth M and therefore smooth $\mathcal{E}$ which is free from self-intersections, to illustrate our approach and its implementation.

Our main contributions are (i) a clear topological description of the sweep, and (ii) an architectural framework for its construction. This, coupled with [6], which constructs the geometry/parametrizations of the surfaces, was used to build a pilot implementation of the solid sweep using the popular ACIS solid modeling kernel [3]. We give several illustrative examples produced by our implementation to demonstrate the effectiveness of our algorithm. To the best of our knowledge, this is the first attempt to explicate the complete brep structure of $\mathcal{E}$.

The solid sweep has been extensively studied [1, 2, 3, 4, 7]. In [2] the envelope is modeled as the solution set of the rank-deficient Jacobian of the sweep map. This method uses symbolic computation and cannot handle general input such as splines. In [3] the authors derive a differential equation whose solution is the envelope. An approximate envelope surface is fitted through the points sampled on the envelope. In [4] the authors give a membership test for a point to belong inside, outside or on the boundary of the swept volume. This does not yield a parametric definition of the envelope. In [5] the trajectory is approximated by a screw motion in order to compute the swept volume. In [7] the evolution speed of the curve of contact is studied in order to achieve a prescribed sampling density of points on the envelope, through which a surface is fit to obtain an approximation to the envelope. For a more comprehensive survey of the previous work, we refer the reader to [1]. Much of the work has focused on the mathematics of the surface. The exact topological structure has not been investigated in any significant detail.

Bibliography

[1] Abdel-Malek K.; Blackmore D.; Joy K: Swept Volumes: Foundations, Perspectives and Applications, International Journal of Shape Modeling. 12(1), 2006, 87-127.

[2] Abdel-Malek K, Yeh HJ. Geometric representation of the swept volume using Jacobian rank-deficiency conditions. CAD 1997;29(6):457-468.

[3] Blackmore D, Leu MC, Wang L. Sweep-envelope differential equation algorithm and its application to NC machining verification. CAD 1997;29(9):629-637.

[4] Huseyin Erdim, Horea T. Ilies. Classifying points for sweeping solids. CAD 2008;40(9);987-998

[5] J. Rossignac, J.J. Kim, S.C. Song, K.C. Suh, C.B. Joung. Boundary of the volume swept by a free-form solid in screw motion. CAD 2007;39; 745-755

[6] Bharat Adsul, Jinesh Machchhar, Milind Sohoni. Local and Global Analysis of Parametric Solid Sweeps. Cornell University Library arXiv. http://arxiv.org/abs/1305.7351

[7] Peternell M, Pottmann H, Steiner T, Zhao H. Swept volumes. CAD and Applications 2005;2;599-608

[8] Seok Won Lee, Andreas Nestler. Complete swept volume generation, Part I: Swept volume of a piecewise C1-continuous cutter at five-axis milling via Gauss map. CAD 2011;43(4);427-441

[9] Seok Won Lee, Andreas Nestler. Complete swept volume generation, Part II: NC simulation of self-penetration via comprehensive analysis of envelope profiles. CAD 2011;43(4);442-456

[10] Kinsley Inc. Timing screw for grouping and turning. https://www.youtube.com/watch?v=LooYoMM5DEo

A Novel Power Model and Completion Time Model for Virtualized Environments

Swetha P.T. Srinivasan Prof. Umesh Bellur

As power consumption costs takes upto half of the operational expenses of a data center, power management is now a critical concern of high performance computing and cloud computing data centers alike. Recent advances in processor technology have provided fine-grained control over the frequency and voltage at which processors operate and increased the dynamic power range of modern computers thus enabling power-performance negotiations. At the same time, the trends towards virtualization of data centers and HPC environments give us a clear boundary for control and helps us map virtual machines (VM) to specific cores of a physical machine. Our ultimate goal is to provision VMs while satisfying the twin and possibly competing requirements of a power budget

and VM performance thresholds. As an intermediate step, with a power and completion time model in place, we could optimize for one or the other of these requirements as well and use the resulting model to understand the effect that maximizing performance will have on power or minimizing power consumption will have on the performance of tasks executing inside VMs. The complexity in effectively providing power management arises from two sources - being able to predict the power consumption at a given performance level accurately and then to use this prediction in VM provisioning. Although many power and performance models exist, the combined effect of processor frequency and compute resource variations on the power usage and performance of VMs are not analyzed thoroughly. Moreover, the models do not perform well for memory-intensive or disk-intensive benchmarks.

In our work, we present an empirically derived power model and a completion time model using linear regression with CPU utilization and operating frequency of the server as parameters. We have validated the power model by using several processors in the Intel 'i'-series, the Xeon series as well as the AMD processor in the x4 series and shown that our model predicts power consumption within a range of 2-7% of the measured power. We have also validated the completion time model by predicting the execution time of six CPU, memory and disk-intensive benchmarks on four heterogeneous systems and our model predicts the completion time within a range of 1-6% of the observed execution time. We are now in the process of employing these models to control VM provisioning to minimize power consumption while at the same time meeting the service level agreements (SLAs) of the applications executing within VMs.

For future work, we intend to extend the power model to dual or n-processor systems, encompass memory, disk and network resources. We will propose a VM provisioning algorithm that is completion time-aware and power optimal across a cloud setup. The algorithm could be extended to include dynamic SLA requirements and reallocation of server resources power optimally.

The Weighted k-server Problem and Randomized Memoryless Algorithms

Ashish Chiplunkar Prof. Sundar Vishwanathan

The weighted k-server problem is a generalization of the k-server problem in which the cost of moving a server of weight β_i through a distance d is $\beta_i \cdot d$. Fiat and Ricklin [4] initiated the study of this problem, motivated by the fact, that the weighted server problem on uniform metric spaces models caching with caches differing in page replacement costs. In our work [2], we prove tight bounds on the performance of randomized memoryless algorithms for this problem. We first prove that there is an α_k-competitive randomized memoryless algorithm for this problem, where $\alpha_k = \alpha_{k-1}^2 + 3\alpha_{k-1} + 1$; $\alpha_1 = 1$. We complement this result by proving that no randomized memoryless algorithm can have a competitive ratio better than α_k. We thus generalize the results of Chrobak and Sgall [3], who gave a 5-competitive randomized memoryless algorithm for the weighted 2-server problem on uniform metrics, and proved that this bound can not be improved.

To prove the upper bound of α_k we develop a framework to bound from above the competitive ratio of any randomized memoryless algorithm for this problem. The key technical contribution is a method for working with potential functions defined implicitly as the solution of a linear system. The result is robust in the sense that a small change in the probabilities used by the algorithm results in a small change in the upper bound on the competitive ratio. The above result has two important implications. Firstly this yields an α_k-competitive memoryless algorithm for the weighted k-server problem on uniform spaces. This is the first competitive algorithm for $k > 2$, which is memoryless. Secondly, this helps us prove that the Harmonic algorithm, which chooses probabilities in inverse proportion to weights, has a competitive ratio of $k\alpha_k$.

The only known competitive algorithm for every k before this work is a carefully crafted deterministic algorithm due to Fiat and Ricklin [4]. Their algorithm uses memory crucially and their

bound on competitive ratio is 2^{4^k}. Our algorithm is not only memoryless, but also has a considerably improved competitive ratio of $\alpha_k < 1.6^{2^k}$. Further, the derandomization technique of Ben-David et al. [1] implies that there exists a deterministic algorithm for this problem with competitive ratio $\alpha_k^2 < 2.56^{2^k}$.

Bibliography

[1] Shai Ben-David and Allan Borodin and Richard M. Karp and Gábor Tardos and Avi Wigderson. On the Power of Randomization in On-Line Algorithms. *Algorithmica*, 11(1):2–14, 1994.

[2] Ashish Chiplunkar and Sundar Vishwanathan. On Randomized Memoryless Algorithms for the Weighted k-Server Problem. In FOCS, pages 11–19. IEEE Computer Society, 2013.

[3] Marek Chrobak and Jiří Sgall. The weighted 2-server problem. *Theoretical Computer Science*, 324(2-3):289–312, 2004.

[4] Amos Fiat and Moty Ricklin. Competitive Algorithms for the Weighted Server Problem. *Theoretical Computer Science*, 130(1):85-99, 1994.

Nonlinear Optimizations for Real-time Systems

Devendra Bhave Prof. Ashutosh Trivedi Prof. Krishna S.

Timed automaton is popular theoretical model for real-time systems. Linearly priced time automata form a class in which waiting cost increases linearly with time. We define class of time automata where prices are permitted to increase non-linearly with time. We show this relaxation, although extremely useful in practice, leads to undecidable results.

We suggest use of integer reset timed automaton (IRTA). This restriction helps us to show decidability for computing optimal reachability problem, without compromising too much on practical usability of our model. We relax conditions on IRTA to define new subclasses of timed automata. We define *extended IRTA* which permits arbitrary clock valuations for initial state. Extended IRTA preserves good properties of IRTA like closure under binary operations and determinizability. We define one more subclass of timed automaton in which integer reset is required along cycles in region graph. We call it *relaxed IRTA* as it relaxes integer reset requirement for acyclic paths. We study optimal cost reachability problem for concave priced relaxed IRTA and prove its PSPACE-completeness.

Bibliography

[1] P. Bouyer, T. Brihaye, V. Bruyere, and JF Raskin. On the optimal reachability problem of weighted timed automata. Formal Methods in System Design, 31(2):135–175, 2007.

[2] M. Jurdzinski and A. Trivedi. Concavely-priced timed automata. In FORMATS, pages 48–62, 2008.

[3] L. Manasa and S. Krishna. Integer reset timed automata: Clock reduction and determinizability. CoRR, abs/1001.1215, 2010.

[4] P. Vijay Suman, P. Pandya, S. Krishna, and Lakshmi Manasa. Timed automata with integer resets: Language inclusion and expressiveness. In FORMATS, pages 78–92, 2008.

Traceability analyses between features and assets in software product lines

Ganesh Khandu Narwane Prof. Shankara Narayanan Krishna A. K. Bhattacharjee

Collaborator(s): David Benavides, Jean-Vivien Millo, S Ramesh, José A. Galindo

In a Software Product Line (SPL), the central notion of *implementability* provides the requisite connection between specifications (feature sets) and their implementations (component sets), leading to the definition of products. While it appears to be a simple extension (to sets) of the traceability relation between components and features, it actually involves several subtle issues which are overlooked in the definitions in existing literature. In this paper, we give a precise and formal definition of implementability over a fairly expressive traceability relation to solve these issues. The consequent definition of products in the given SPL naturally entails a set of useful analysis problems that are either refinements of known problems, or are completely novel. We also propose a new approach to solve these analysis problems by encoding them as Quantified Boolean Formula(QBF) and solving them through Quantified Satisfiability (QSAT) solvers. QBF can represent more complex analysis operation, which cannot be represented by using propositional formulae. The methodology scales much better than the SAT-based solutions hinted in the literature and is demonstrated through a tool called SPLAnE (SPL Analysis Engine), on a couple of fairly large case studies.

Bibliography

[1] Swarup Mohalik and S. Ramesh andJean-Vivien Millo and Shankara Narayanan Krishna and Ganesh Khandu Narwane. Tracing SPLs precisely and efficiently. 16th International Software Product Line Conference, SPLC '12, Salvador, Brazil - September 2-7, 2012, Volume 1.

[2] http://www.cse.iitb.ac.in/ krishnas/splane/

A Convex Feature Learning Formulation for Latent Task Structure Discovery

Pratik Jawanpuria Prof. Saketha Nath

Multi-task learning (MTL) paradigm advocates joint training of several *related* learning problems (tasks). The rationale behind MTL is that by *intelligently* sharing the features learnt during the training phase, the generalization across all the tasks may get better (compared to learning them independently). This gives rise to two important concerns that needs to be taken care of while performing MTL: a) *which* features to share, and b) with *whom* to share. Both sharing unhelpful feature information among related tasks or sharing any information with unrelated tasks may result in worsening of performance. The focus of this work revolves around both these issues.

We consider a setting where some relevant features could be shared across few related tasks. Most of the existing methods assume the extent to which the given tasks are related or share a common feature space to be known apriori. In real-world applications however, it is desirable to automatically discover the groups of related tasks that share a feature space. In this work we aim at searching the exponentially large space of all possible groups of tasks that may share a feature space. The main contribution is a convex formulation that employs a graph-based regularizer and simultaneously discovers few groups of related tasks, having close-by task parameters, as well as the feature space shared within each group. The regularizer encodes an important structure among the groups of tasks leading to an efficient algorithm for solving it: if there is no feature space under which a group of tasks has close-by task parameters, then there does not exist such a feature space for any of its supersets. An efficient active set algorithm that exploits this simplification and performs a clever search in the exponentially large space is presented. The algorithm is guaranteed to solve the proposed formulation (within some precision) in a time polynomial in the number of groups of related tasks discovered. Empirical results on benchmark datasets show that the proposed

formulation achieves good generalization and outperforms state-of-the-art multi-task learning algorithms in some cases.

Bibliography

[1] @ONLINE http://www.cse.iitb.ac.in/ pratik.j/research.html

Towards Personalization of Sentiment Analysis

Aditya Joshi Prof. Pushpak Bhattacharyya Prof. Mark J Carman

Collaborator(s): Abhijit Mishra, Nivvedan Senthamilselvan

Sentiment Analysis (SA) is the task of automatic prediction of opinion in a textual unit (which may be a user review, tweet, etc.). However, the user-side component of sentiment analysis has seldom been considered. For example, the news article '*Leader X arrested in Mumbai today*' arguably carries no sentiment of its own. However, based on the political orientation of a reader, the article may evoke positive or negative sentiment in the reader. While traditional sentiment analysis deals with sentiment expressed by the writer of a textual unit, we aim to explore how sentiment evoked in a reader may be predicted. Towards this, we explore two directions: (A) cognitive understanding of sentiment analysis in humans, and (B) estimation of topic models for sentiment.

Our cognitive study uses an eye-tracking device to understand how readers identify sentiment in a textual unit. The first study records eye movements of readers that have been assigned the task of predicting sentiment in three reviews. We observe two distinct behaviours: anticipation (where users skip sentences) and homing (where users read the document and then home back to a subset of sentences before making a judgment). Which of the two options is chosen depends on the nature of sentiment within a document. This interplay of eye-tracking features and linguistic aspects of a

document is then exploited in our second study. This study aims to automatically detect how difficult a document is for a sentiment annotator. We obtain a data set of 1059 movie review snippets and obtain annotation from five human participants. This annotation is in the form of eye-tracking data. We then learn a classifier that uses linguistic attributes as features, average fixation duration as the label and predicts a score called ' textitsentiment annotation complexity (SAC)'. Such a score allows test documents to be binned based on their sentiment annotation complexity and may be used to refine crowd-sourcing pricing models.

The second part of our work focuses on *probabilistic topic models* for sentiment. Topic models are a suite of algorithms that discover thematic structures underlying a data collection. We present primary experimental results to extract sentiment word-lists from a data collect. We implement two known classes of LDA models: the first uses sentiment priors using sentiment lexicon while the second adds sentiment as additional variables. We propose two future directions: generation of user-specific sentiment models and validation of structure of basic emotions using correlated topic models.

Knowledge Representation Approaches for Information Retrieval in Hindustani Music

Joe Cheri Ross Prof. Pushpak Bhattacharyya Prof. Preeti Rao

Efficient information management is critical in information retrieval for music. Since the existing ontology design for knowledge representation is insufficient to cater the requirements of knowledge representation in Indian music, an ontology has to be designed with provision to represent terms and concepts pertaining to Indian music. The need of good representation arose from its strong relation with information retrieval. Gopal et. al has proposed design of an ontology for Indian music emphasizing on Carnatic music [2]. This design enables representation of concepts

like raga, *svara*, *gamaka* making use of sequence design patterns. This has to be extended for representation of concepts in Hindustani music and other generic concepts including phrases(motifs), *tala* etc. Design of ontology has to be driven by identification of genre specific concepts and generic concepts within Indian music.

To enable better information retrieval for music along with representation content based information, music meta data in textual form are also to be a part of music ontology. This work focussing on information extraction from online sources (rasikas.org, wikipedia) has two primary modules: one to extract the relevant information, such as entities, their properties and relationships from each of the aforementioned sources, and another module to map and interlink such information using our ontologies which will result in a multimodal knowledge-base. Natural language processing (NLP) based approach is adopted for information extraction, analysing syntactic patterns present in the content. This work is focussing on a dictionary based named entity identification approach and methods to resolve coreferences.

Bibliography

[1] @ONLINE http://www.rasikas.org

[2] Gopala Krishna Koduri and Xavier Serra. Knowledge-based approach to computa- tional analysis of melody in indian art music. In First International Workshop on Semantic Music and Media, 2013.

Harnessing Annotation Process Data for NLP-An Investigation based on EYE-TRACKING

Abhijit Mishra Prof. Pushpak Bhattacharyya

Collaborator(s): Prof. Michael Carl

The state-of-the-art NLP techniques deal with *Machine Learning(ML)* and *Linguistics*. Over the years, supervised and semi-supervised ML methods have been extensively used to design high quality systems for NLP tasks like *Part of Speech Tagging, Parsing, Disambiguation* and *Machine Translation*. The objective of these techniques is to find mathematical models that best explain large scale human/machine annotated data. The system based on these techniques are classified as weak *Artificial Intelligence* (AI) systems and can hardly produce output beyond the scope of the training data. On the other hand, humans do not need large number of examples to learn a language phenomenon. So, it is desirable to build NLP systems that perform tasks in the same manner with similar accuracy as humans. These systems would be called as *strong-AI* systems. In order to achieve as good a system as strong-AI systems, a detailed understanding of the human way of language processing is necessary. Hence, in our work, instead of proposing new weak-AI systems, we have taken a step back to capture and analyze the activities during annotation in order to explain the cognitive processes underlying human language processing. Such insights will be translated to better algorithms in NLP.

Our aim is to exploit the process of manual text-annotation by capturing the "eye-gaze behavior" of the annotators. We classify this captured data as *Annotation Process Data (APD)*. APD comprises *gaze-fixations, saccades, eye-regression patterns* and *scan-paths*.

It is conceivable that the eye-movement patterns will vary based on certain linguistic peculiarities of the text being annotated. Through our eye-tracking experiments, we try to find out

these specific linguistic properties and use them in Machine Learning settings. Our initial attempts try to address two different problems in NLP *viz.* 1. *Predicting Sentence Translation Difficulty* 2. *Using Human Subjectivity Extraction strategies towards improving Sentiment Analysis algorithms.* We also enlist several possibility to utilize APD for *Word Sense Disambiguation* and *Parsing.* Our overall observation is that, harnessing behavioral data looks promising for modern NLP.

Bibliography

[1] Abhijit Mishra, Pushpak Bhattacharyya and Michael Carl.2013. Automatically Predicting Sentence Translation Difficulty. *ACL 2013*, Sofia, Bulgaria

[2] Aditya Joshi, Abhijit Mishra, Pushpak Bhattacharyya. 2013. A cognitive study of subjectivity extraction in sentiment annotation. *Under review*

Explorations in Statistical Machine Translation for Indian Languages

Anoop Kunchukuttan Abhijit Mishra Prof. Pushpak Bhattacharyya

Collaborator(s): Rajen Chatterjee, Ritesh Shah, Shourya Roy

We present the following work related to SMT systems for Indian languages:

***Śata-Anuvādak*: Tackling Multiway Translation of Indian Languages [1]**

We present a compendium of 110 Statistical Machine Translation systems built from parallel corpora of 11 Indian languages belonging to both Indo-Aryan and Dravidian families. We attempt an analysis of the relationship between translation accuracy and the language families involved. Insights obtained therefrom,

we feel, will provide guidelines for creating machine translation systems of specific Indian language pairs. We build phrase based systems and some extensions. Across multiple languages, we show improvements on the baseline phrase based systems using these extensions: (1) source side reordering for English-Indian language translation, and (2) transliteration of untranslated words for Indian language-Indian language translation.

URL: http://www.cfilt.iitb.ac.in/indic-translator

Map-Reduce based Crowdsourced Translation for Complex Domains [2]

We present the *TransDoop* system for gathering translations to create parallel corpora from online crowd workforce who have familiarity with multiple languages but are not expert translators. Our system uses a Map-Reduce-like approach to translation crowdsourcing where sentence translation is decomposed into the following smaller tasks: (a) translation of constituent phrases of the sentence; (b) validation of quality of the phrase translations; and (c) composition of complete sentence translations from phrase translations. For a complex domain like judicial proceedings, the higher scores obtained by the map-reduce based approach compared to complete sentence translation establishes the efficacy of our work.

Bibliography

[1] Anoop Kunchukuttan, Abhijit Mishra, Rajen Chatterjee, Ritesh Shah and Pushpak Bhattacharyya. Śata-Anuvādak: Tackling Multiway Translation of Indian Languages. Language Resources and Evaluation Conference. 2014.

[2] Anoop Kunchukuttan, Rajen Chatterjee, Shourya Roy, Abhijit Mishra and Pushpak Bhattacharyya. TransDoop: A Map-Reduce based Crowdsourced Translation for Complex Domain . Proceedings of ACL. 2013.

Detecting granularity in Words: for the purpose of Sentiment Analysis

Raksha Sharma Prof. Pushpak Bhattacharyya

Sentiment Analysis (SA) has grown tremendously over the decade. More and more sophisticated techniques are built to tackle the problem. The evolution of methods has been on several different dimensions. Some of these are complexity of algorithm, the knowledge source used, *etc.* The SA task is to predict the sentiment orientation of a text (document/para/sentence) by analyzing the polarity of words present in the text. A lexicon of sentiment bearing words is of great help in such tasks.

The research however has failed to tackle some of the rare yet important problems in sentiment analysis. One such problem is ***granularity in polar words***. Mere identification of a word as a polar word is inadequate for deep sentiment analysis. For example, ***unpredictable*** is positive in the movie domain, but negative in the product domain. We call these words *chameleon words.* Identification of such ***domain dependence*** in polar words is a pivotal task for domain-dependent sentiment analysis.

On the other hand, there exists ***domain specific words***, which have the same polarity across domains, but are used very frequently in a particular domain. For example, *miscast, blockbuster*, are used very frequently in the movie domain. Chameleon words and domain specific words, can be united as ***Domain Dedicated Polar Words (DDPW)***.

One more dimension of granularity in polar words is ***Polarity-intensity variation across synonyms***. For example, the synset {*sound, levelheaded, level-headed, intelligent, healthy*} (Gloss: exercising or showing good judgment), has a positive polarity of *0.75* in SentiWordNet, while most people would agree that all the words in the synset are not equally positively intense for the given sense. The use of *levelheaded* or *sound* makes a sentence more intensely positive in comparison to *healthy*, given that the sentence expresses the sense *exercising or showing good judgment.* Polarity-intensity variation across synonyms is a milestone to improve on

existing sentiment scores in sentimental-WordNets, consequently the improvement on state of art sentiment analysis system.

These granularity issues with the polar words, though recognized by a few researchers are still not approached by any researcher. Our work mainly tackles these problems. Through this work we describe our understanding of the problem and also a way to solve these problems using the intuitions that govern human decisions when presented with this phenomenon. The aim of this work was to build a system that detects ***Domain-dedicated polar words*** and ***Polarity-intensity variation across synonyms***. Such a system can be used to improve the performance of a sentiment analysis system.

Semi-supervised Relation Extraction using EM Algorithm

Sachin Pawar Prof. Pushpak Bhattacharyya Girish Palshikar

Relation Extraction is the task of identifying relation between entities in a natural language sentence. We propose a semi-supervised approach for relation extraction based on EM algorithm, which uses few relation labeled seed examples and large number of unlabeled examples (but labeled with entities). We present analysis of how unlabeled data helps in improving the overall accuracy compared to the baseline system using only labeled data. This work therefore shows the efficacy of a sound theoretical framework exploiting an easily obtainable resource named "unlabeled data" for the problem of relation extraction.

Our technique is motivated from the semi-supervised text classification method proposed by Nigam et.al. [1]. We modelled relation extraction problem using a discriminative EM algorithm instead of the generative EM based technique used by Nigam et.al. For capturing characteristics of various relations, it is quite difficult to design features which are non-overlapping and independent, as required by the generative models. Hence we chose

to model relation extraction problem using discriminative Maximum Entropy classifier in the EM setting. We test our technique on two different corpora - CoNLL 2003 NER shared task corpus and Agriculture news corpus. The CoNLL corpus is labeled with Named Entities - PERSON, ORGANIZATION and LOCATION and the relations as described in Doddington et.al. [2] were considered for extraction. In the agriculture news corpus, we considered CROP and DISEASE named entities and the relation *Affects* between them. We demonstrate effectiveness of our technique by achieving higher performance than two baselines - first one using only labeled data and the second baseline using Co-Training algorithm [3].

Bibliography

[1] Kamal Nigam, AK McCallum, S Thrun, T Mitchell 2000, Text Classification from Labeled and Unlabeled Documents using EM , Machine Learning, 39, 103-134.

[2] George Doddington, Alexis Mitchell, Mark Przybocki, Lance Ramshaw, Stephanie Strassel, Ralph Weischedel 2004, The Automatic Content Extraction (ACE) Program Tasks, Data, and Evaluation LREC 2004.

[3] Blum, Avrim, and Tom Mitchell 1998, Combining labeled and unlabeled data with co-training, Proceedings of the eleventh annual conference on Computational learning theory. ACM, 1998.

Unsupervised Word Sense Disambiguation

Sudha Bhingardive Prof. Pushpak Bhattacharyya

Collaborator(s): Samiulla Shaikh

Word Sense Disambiguation (WSD) is one of the toughest problems in NLP. Almost all applications in NLP are directly or indirectly related to WSD *viz.,* Sentiment Analysis, Machine Translation, Text summarization, Text Entailment, Semantic Role Labeling, etc. Successful supervised WSD approaches are restricted

to resource rich languages and domains. They are directly dependent on availability of good amount of sense tagged data. Creating such a costly resource for all language-domain pairs is impracticable looking at the amount of time and money required. This situation demands the development of approaches which use the minimal amount of these resources. Hence, unsupervised WSD approaches attract most of the researchers.

In WSD, verb disambiguation has proved to be extremely difficult, because of high degree of polysemy, too fine grained senses, absence of deep verb hierarchy and low inter annotator agreement in verb sense annotation. Unsupervised WSD has received widespread attention, but has performed poorly, specially on verbs. Recently an unsupervised bilingual EM based algorithm [1] has been proposed, which makes use only of the raw counts of the translations in comparable corpora (Marathi and Hindi). But the performance of this approach is poor on verbs with accuracy level at 25-38%.

We suggest a modification to this mentioned formulation, using context and semantic relatedness of neighboring words [2]. An improvement of 17% - 35% in the accuracy of verb WSD is obtained compared to the existing EM based approach. On a general note, the work can be looked upon as contributing to the framework of unsupervised WSD through context aware expectation maximization. Future directions point to usage of semantic role clues, investigation of familialy apart pair of languages and effect of variation of measures of semantic relatedness.

Bibliography

[1] Mitesh Khapra, Salil Joshi and Pushpak Bhattacharyya, "It Takes Two to Tango: A Bilingual Unsupervised Approach for Estimating Sense Distributions using Expectation Maximization", IJCNLP, Chiang Mai, Thailand, 2011.

[2] Sudha Bhingardive, Samiulla Shaikh and Pushpak Bhattacharyya, "Neighbors Help: Bilingual Unsupervised WSD Using Context", ACL, Sofia, Bulgaria, 2013.

Derivation of Imperative Programs from Formal Specifications

Dipak L. Chaudhari Prof. Om Damani

The prevalent *implement-and-verify* software development methodologies typically involve an implementation stage and a separate verification stage. Testing/Formal verification is usually employed to gain confidence that the implemented programs are defect-free. Although the failed test cases or the failed proof obligations provide some hint, there is no structured help available to the user in the actual task of implementing the programs. Programmers rely on ad-hoc use cases and informal reasoning to guess the program constructs.

Calculational Style of Programming is a correct-by-construction programming methodology wherein programs and the correctness proofs are systematically derived together from the formal specifications. Program constructs and the associated invariants are naturally discovered during the program derivation process. Many aspects of this method, however, are mechanical and error prone if done manually. Our aim has been to build a program synthesis platform wherein theorem provers assist the user in performing the mundane formula manipulation tasks taking the drudgery out of the derivation process. Guessing the program constructs and associated invariants is a challenging activity that needs creativity and application of algorithm design techniques. Our objective has been to enrich the program synthesis platform by capturing the algorithm design techniques in the form of synthesis tactics.

Tools like *Why3*, *Spec#*, and *Dafny* try to bring the verification stage closer to the implementation stage. These tools provide limited support to the programmer in the task of implementing the programs but fail to capture the rationale behind the introduction of program constructs and invariants.

We have designed and implemented an interactive tool, *ProgSynth*, for deriving programs from formal specifications. Modeling the back and forth iterations in the derivation is crucial since it captures the rationale behind the design choices. The

tool maintains the complete derivation history in the form a Synthesis Tree. User can backtrack to any point of the derivation and branch off to explore different derivation strategies. To simplify the derivation, ProgSynth allows transformations on focused subcomponents. Additional contextual information is made available to the user which can be used for manipulating the subcomponent. We have designed more powerful transformation rules that employ an external theorem prover (SMT solver, Z3) in addition to using the prevalent rewrite based techniques. We have captured commonly used program derivation principles in the form of synthesis tactics. ProgSynth enables user to focus on the creative aspect of the program derivation process by taking care of underlying details

Bibliography

[1] Rajeev Alur, Rastislav Bodik, Garvit Juniwal, Milo M. K. Martin, Mukund Raghothaman, Sanjit A. Seshia, Rishabh Singh, Armando Solar-Lezama, Emina Torlak, and Abhishek Udupa. Syntax-guided synthesis. In Proceedings of the IEEE International Conference on Formal Methods in Computer-Aided Design (FMCAD), October 2013.

[2] Armando Solar-Lezama, Liviu Tancau, Rastislav Bodik, Sanjit Seshia, and Vijay Saraswat. Combinatorial sketching for finite programs. ACM SIGARCH Computer Architecture News, 34(5):404-415, 2006.

[3] Edsger Wybe Dijkstra. A Discipline of Programming. Prentice Hall PTR, Upper Saddle River, NJ, USA, 1997.

[4] Anne Kaldewaij. Programming: the derivation of algorithms. Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 1990.

[5] Ralph J. Back and Joakim Wright. Refinement Calculus: A Systematic Introduction (Texts in Computer Science). Springer, April 1998.

On Satisfiability of Metric Temporal Logic

Khushraj Madnani Prof. Shankara Narayanan Krishna
Prof. Paritosh K. Pandya

Metric Temporal Logic, $\mathsf{MTL}^{pw}[\mathcal{U}_I, \mathcal{S}_I]$, is amongst the most studied real-time logics. It exhibits considerable diversity in expressiveness and decidability properties based on the permitted set of modalities and the nature of time interval constraints I. In this paper, we sharpen the decidability results by showing that the satisfiability of $\mathsf{MTL}^{pw}[\mathcal{U}_I, \mathcal{S}_{NS}]$ (where NS denotes non-singular intervals) is also decidable over finite point-wise time. We give satisfiability preserving reduction from $\mathsf{MTL}^{pw}[\mathcal{U}_I, \mathcal{S}_{NS}]$ to the decidable logic $\mathsf{MTL}^{pw}[\mathcal{U}_I]$ of Ouaknine and Worrell using technique of temporal projections [1] [3] [2] [4]. We generalize the technique for point-wise semantics with 2 variations (i) where we do not allow any addition of extra time pointsc (ii) where the tranformed behaviours is oversampling of original behaviour(i.e. allowing extra time points) using a novel technique of temporal projections with oversampling. The reduction using (i) is not only complex but also has a exponential(pseudo-polynomial) blow up in the formula size. Oversampling technique here is giving us equisatisfiable formula in $\mathsf{MTL}^{pw}[\mathcal{U}_I]$ which is linear with respect to initial formula in $\mathsf{MTL}^{pw}[\mathcal{U}_I, \mathcal{S}_{NS}]$.

We also investigate the decidability of unary fragment $\mathsf{MTL}^{pw}[\mathcal{F}_I, \mathcal{P}_I]$.

Bibliography

[1] Deepak D'Souza and M Raj Mohan and Pavithra Prabhakar. Eliminating past operators in Metric Temporal Logic. *Perspectives in Concurrency*, 86–106, 2008.

[2] Y. Hirshfeld and A. Rabinovich. Logics for Real Time: Decidability and Complexity. *Fundam. Inform.*, 62(1), 2004, 1-28.

[3] D.Kini, S. N. Krishna and P. K.Pandya. On Construction of Safety Signal Automata for $MITL[\mathcal{U}_I, \mathcal{S}_I]$ using Temporal Projections. *Proceedings of FORMATS 2011*, 225-239.

[4] P. Prabhakar and Deepak D'Souza. On the Expressiveness of MTL with Past Operators. *Proceedings of FORMATS 2006*, 322–336.