ACCELERATE BC

BC's Graduate Research Internship Program

Final Report

Research Areas (Specific keywords):

Title of Project: Probabilistic Description, Modeling and Similarity-Ranking of Mineral Deposit Zones.

Period of Internship (start and end date)**:** January 1, 2011 – April 30, 2011

Date of Submission: June 8, 2011

Participants

Intern:

Name: Jacek Kisyński
Degree program: PDF
University: University of British Columbia
Department: Computer Science
Address: 201-2366 Main Mall, Vancouver, BC V6T 1Z4 Canada
Phone: 604-822-3061
Fax: 604-822-5485
Email: kisynski@cs.ubc.ca
Male/Female: male
Citizenship: Polish

Supervisor:

Name: Giuseppe Carenini
University: University of British Columbia
Department: Computer Science
Address: 129-2366 Main Mall, Vancouver, BC V6T 1Z4 Canada
Phone: 604-822-5109
Fax: 604-822-5485
Email: carenini@cs.ubc.ca

Organization Sponsor:

Name:	Steve Williams	Clinton Smyth
Position:	President	President
Organization:	GMT GeoMetTech Ltd.	Georeference Online Ltd.
Address:	1441 Murphy Rd, Ennismore, ON 301-850	West Hastings Street, Vancouver, BC V6C 1E1
Phone:	705-292-8116	604-692-0850
Fax:		604-692-0850
Email:	stevewilliams@geomettech.com	cpsmyth@msn.com

REPORT

1. Background information.

Georeference Online Ltd (GOL) has developed a matching system which has been applied to minerals exploration, geohazards mapping and metallurgical processing. The system uses ontologies to describe instances (e.g. specific zones in specific mineral deposits) and models (e.g. archetypal deposit zones which are associated with optimal metallurgical processing). It is then able to compare descriptions, rank them based on similarity, and explain to the user how the rankings have been derived.

GOL's matching system (see Poole et al. 2005) used to use, with a considerable degree of success, additive costs, which are informal log-probabilities, to measure similarity. In a study partially funded by MITACS in 2006, GOL re-examined the mathematical fundamentals of its reasoning with a view to providing a more rigorous method of dealing with the prior probabilities. The study placed the additive cost approach on firmer foundations (see Sharma et al. 2010). GOL also addressed the role of taxonomies in the matching and the critical role of Aristotelian approach to class definitions in these taxonomies (see Poole et al. 2009).

In 2010, during two months of privately funded research and four months of MITACS-supported research, GOL has attempted to implement the results of Poole et al. 2009 and Sharma et al. 2010 into operational software written in Java, but have not completed the work because, during its course, we uncovered issues and complexities that warranted further investigation. This MITACS Accelerate BC internship aimed to continue research described above.

2. What were the research goals?

1. Adding number-line reasoning to the probabilistic matching algorithm based on the work for an early version of the matching system (see Huang et al. 2005).
2. Investigate improving the current Matcher with an alternative that uses noisy-and and noisy-or to represent probability of a model given the description.
3. Investigate incorporating the utility of the models as well as the probabilities into their similarity rankings.
4. Implementation and empirical evaluation of the updated matching system including interfacing to user-interface code to be written by a Georeference Online Ltd software engineer.

3. Goals achieved.

Goal 1 has been partially achieved. Goal 2 has been achieved but with a different outcome than predicted. Instead of using noisy-and and noisy-or, the Matcher has been improved by adopting calculus of odds. Goal 3 has been abandoned as the other goals proved to be more important. Goal 4 has been achieved.

Tasks that were not described in the internship proposal, but were addressed during the internship:

- Matcher meta-ontology and the matching algorithm have been extended to account for functional dependencies between properties present in the input ontologies. As a result quality of scores returned by the Matcher improved dramatically.

4. What were the methods or techniques applied?

The matcher software is being implemented in the Java programming language. Development is carried out inside the Eclipse IDE. Accompanying ontologies are specified in OWL (see W3C 2009) and created using TopQuadrant's TopBraid Composer ontology editor. Matcher code interacts with ontologies via OWL

API library (see Horridge et al. 2009). Descriptions of the deposits (both models and instances) are stored either in RDF/XML files (see W3C 2004) or in a triplestore. We use Franz's AllegroGraph RDFStore. Matcher uses the JUnit framework for unit testing and SL4J / Logback frameworks for logging.

5. Summary of the outcome of the research.

Calculus of odds proved to be much more suitable for the Matcher than operations on conditional probability tables. Accounting for functional dependencies between properties resulted in score quality improvements. Detailed score explanations displayed in user-interface created by a Georeference Online Ltd software engineer allowed iterative improvement of the Matcher based on the feedback from domain experts.

6. What are the benefits to the Partner Organization from this Internship?

The internship gave GOL a functioning matching system based on firm theoretical foundations. Further enhancements to the systems can be test-driven and based on the feedback from domain experts.

GMT GeoMetTech Ltd. expects to apply the matching system to geometallurgy by more efficiently recognizing similarities between geological zones than would be possible by human means alone.

The intern continues to collaborate with GOL beyond the period of the internship.

7. What are the future research plans?

Further explore probabilistic representations different from the currently used naive Bayes representation. Investigate adding utilities to the matching system to allow for explicit expected utility trade-offs.

8. References

Horridge et al. 2009 - Matthew Horridge, Sean Bechhofer, *The OWL API: A Java API for Working with OWL 2 Ontologies*, OWLED 2009, 6th OWL Experienced and Directions Workshop, Chantilly, Virginia, 2009.

Huang et al. 2005 - Erica Huang, Clinton Smyth and David Poole, *Representing and Reasoning with Intervals*, 9th Annual Conference of the International Association for Mathematical Geology, Toronto, Canada, 2005.

Poole et al. 2005 - David Poole and Clinton Smyth, *Type Uncertainty in Ontologically-Grounded Qualitative Probabilistic Matching*, 8th European Conference on Symbolic and Qualitative Reasoning About Uncertainty, Barcelona, Spain, 2005.

Poole et al. 2008 - David Poole, Clinton Smyth and Rita Sharma, Semantic Science: Ontologies, Data and Probabilistic Theories, *Lecture Notes in Computer Science*, 5327:26-40, 2008.

Poole et al. 2009 - David Poole, Clinton Smyth and Rita Sharma, Ontology Design for Scientific Theories That Make Probabilistic Predictions, *IEEE Intelligent Systems*, 24(1):27-36, 2009.

Sharma et al. 2010 - Rita Sharma, David Poole and Clinton Smyth, A framework for ontologically-grounded probabilistic matching, *International Journal of Approximate Reasoning*, 51(2):240-262, 2010.

W3C 2004 - RDF/XML Syntax Specification (Revised), W3C Recommendation, 10 February 2004

W3C 2009 - OWL 2 Web Ontology Language Document Overview, W3C Recommendation, 27 October 2009.