MITACS Accelerate Application

Title: Clustering methodology for thought networks

Period of Internship: May 1, 2010 – April 30, 2011
Date of Submission: March 23, 2010

Is the intern and/or supervisor currently working on a MITACS NCE project? *No*

Is this internship related to a MITACS Consortium member? *Yes, National Institute for Complex Data*

Intern:

Name: Wu Zhou
Degree program: PDF
University & Department: University of Waterloo,
Statistics and Actuarial Science
Address (at university): 200 University Avenue West
City, Province: Waterloo, Ontario
Postal code: N2L 3G1
Phone: 519 888 4567 ext 35999
Fax: 519 746 1875
Email: wzhou@uwaterloo.ca
Alternate Email:

Citizenship: Canadian
Personal:

- Sex: Male
- Francophone: __Yes x_ No
- Aboriginal: __Yes x_ No
- Person with a disability: __Yes x_ No
- First in your family to attend university: x_Yes __ No

Academic supervisor:

Name: Professor R. Wayne Oldford
University & Department: University of Waterloo,
Centre for Computational Mathematics in Industry and Commerce

Address: Centre for Computational Mathematics in Industry and Commerce
University of Waterloo
200 University Avenue West
City, Province: Waterloo, Ontario
Postal code: N2L 3G1
Phone: 519 888 4567 Ext 35999
Fax: 519 888 4313
Email: rwoldford@uwaterloo.ca

Organization Sponsor:

Organization Name: Primal Fusion Inc.
Contact Name: Dr. Ihab Ilyas
Position: Vice-President of Research
Address: 7-258 King Street North
City, Province: Waterloo, Ontario
Postal code: N2J 2Y9
Phone: 519 741 1243 Ext. 204
Email: ihab.ilyas@primalfusion.com

The proposal

1. Background information

Primal Fusion works in the area of consumer-directed creation of semantic networks. A motivating conceptual model is to imagine posing the question "What are you thinking?" and (whatever the user's answer) to be able to deliver a rich semantic network of relevant concepts ranging from those the user will recognize as strongly related to those which might only be loosely related, or surprising, even thought provoking. The user receives a network of associated concepts, or thoughts, to explore and to further develop and organize. The result is a rich and personal organization of information resources. Further, as these semantics are machine readable, they enable a host of personalization and automation applications for users (e.g. searching for information, creating documents and reports, collaborating across social networks).

More concretely, the goal is to provide a robust conceptual and computational framework that will facilitate consumer discovery and development of such personal semantic networks. The networks are informed by, and connected to, a variety of semantically rich and reliable sources (e.g. large ontologies). These are pre-processed in staging and analysis phases (below) that capture essential information on concepts and relations (with some intentional loss to encourage novel associations). The resulting framework provides common structure for diverse and changing source information – it must be possible for example to update this structure regularly over time. The user's interests are matched against this structured information to create a new personal semantic network ("synthesis") and to present it to the user (e.g. visualized as a "thought cloud"). The user may then manipulate this network, editing it, changing its emphasis, and growing it by further interaction with the pre-processed information (i.e. via further synthesis interaction). This semantic network becomes an important personal information resource. Moreover, as users create their own semantic networks, it is intended that these networks also be eligible for harvesting of semantically rich information potentially to be incorporated into the system's aggregate semantic base, adding concepts, relations and strengthening connections.

The company's platform includes different technologies of semantic representation and networking, using both analytical and synthetic approaches. Primal Fusion also designs and develops applications that automate user tasks, interacting with other services on the Internet. A key aspect of Primal Fusion's work is to provide the ability to manage rapidly growing data resources in an economical and manageable way, and to leverage machine automation to create breakthrough products for both producers and consumers of digital media. Brief descriptions of some of the key platform components and interactions follow.

Staging

Primal Fusion extracts, transforms, and loads content into the system for semantic analysis.

Analysis

The analysis engine deconstructs all the staged information to an elemental level, extracting semantic data and compressing it into a dynamically updated knowledge base comprising a large reference-level semantic network. Key to this network is the characteristic of semantic distance between nodes. This post-doc will have the goal of advancing the state of the art in manipulating semantic networks in order to identify and cluster nodes that are related in meaning.

Synthesis

To choose and synthesize concepts, the system uses analyzed concepts together with context information that users provide. Indirect sources, such as existing documents, can also provide context information.

User concept definitions and output

The application directs newly synthesized concepts to the user; for example, as part of a knowledge base, search results, semantically synthesized documents, a web page, or any other customized application taking user input and displaying a result.

Complex-Adaptive Feedback

The reference semantic network is dynamically updated based on new sources of content (for example, from updates to Wikipedia, WordNet, Open Linked Data, and the like). This input data will need to undergo semantic clustering. Additionally, we will want to use similar methods to update the network based on user interactions, including user additions of new concepts (nodes). It is expected that similar approaches can be used for both purposes.

Applications

Primal Fusion applications display consumer-created semantic networks using a variety of visualization techniques such as tag clouds (called “thought clouds” when they represent user ideas and interests). By adding semantic weight or clustering to these clouds, we can improve the overall user experience. This post-doc project may be extended to using clustering for data visualization.

2. Research project proposal

Project objectives

The proposed research will develop formal mathematical representations for the structure of the problems addressed by Primal Fusion’s methodology. Existing Primal Fusion methods will be expressed in terms of this formalism to better understand their strengths and weaknesses and to lead to the development of novel well-founded and scalable methods.

Uncertainty is implicit in the problem Primal Fusion has undertaken – in concept formation, relations between concepts, and in the putative strength of either – and an important part of any proper methodological approach adopted. Appropriate probability models will be developed and applied

where possible to provide natural and coherent means of updating each uncertainty in light of new data – i.e. via Bayes theorem. Related scientific objectives are how best to adapt statistical methods such as cluster analysis to the changing domain presented by multiple semantic network sources and by dynamically updating networks.

More detailed research objectives include the study of the appropriateness of well-known measures of semantic similarity as they apply in this context and the possible development of new methods motivated by probabilistic reasoning and/or by statistical combination of competing similarity measures. Different similarities will produce different clusters via the same method. Alternatively, Wu Zhou's thesis provides a theoretical framework that allows clustering outcomes to be objectively evaluated in experimental conditions and, perhaps more importantly here, shows how to coherently combine different clustering outcomes however arrived at. The scientific merits of these approaches (e.g. combining similarities and clustering, clustering on different similarities and structures and then combining) will be investigated in the context of Primal Fusion's various semantic networks. How best to update these (viz. similarities, clusters, etc.) will also be an important consideration.

Proposed approach and methods

Information sources processed in analysis can be thought of as different ontologies (e.g. *WordNet*, *Wikipedia*, …), first, in the sense that each is a formal representation of concepts and relations between them and, second, in that the ontologies may have differing concepts and relations from one another. Some concepts are simple, arguably atomic, representations – as in the case of a single word. Other concepts such as document section headers are complex concepts, consisting of atomic concepts in some composition (e.g. as conjunction). Further, each of these concepts (simple or complex) can be related to one another in many different ways (e.g. "is-a", "part-of", "synonym", "sub-heading", etc.).

The appropriate abstraction is that of a directed acyclic graph whose nodes represent concepts and whose edges represent relations of a variety of types. More abstractly, at the beginning of analysis we have a collection of nodes N_1, N_2, … , N_n arranged in a DAG. Each node has a concept attached to it, say $C_i = C(N_i)$. Further, each such concept consists of one or more constituent concepts. Concepts are either *complex* or *simple* with the majority of nodes in the initial DAG corresponding to complex concepts. For example, from the *Wikipedia* ontology, the concept "Canadian Football League" might be represented as a *complex concept* having constituents "Canadian Football League", "Canadian Football", "Football League", "Canadian", "Football", and "League". A parent of this node would be the node representing the complex concept "Sports Leagues in Canada" and a child node would be "History" (being a section header). "History" is a *simple concept* – a concept that has only itself as a constituent concept is a *simple concept*, otherwise it is a *complex concept*. Some important goals of analysis include" relating "Sports" to "History", to create new complex concepts e.g. "Canadian Football History", relate "ball" (appearing elsewhere in the DAG) and "sports" to "Canadian Football".

The proposed research will focus first on the above abstraction as given and look toward how best to make reliable inferences about relating concepts to one another within this abstraction. The first item to be explored more thoroughly is that of determining appropriate measures of semantic similarity.

Two standard approaches immediately present themselves. The first is that of Tversky's (1977) *feature matching*. This approach measures the similarity of two sets using the number of objects they share compared to the number in which they differ and so it is a natural measure of similarity between *complex* concepts. The second is that of path distance which measures dissimilarity between nodes in a graph by the shortest path between them. This too is natural for complex concepts though it will certainly differ from the Tversky (1977) similarity. Neither method provides an obvious way of relating concepts which appear as constituents of different complex concepts.

Developing such a measure will be one of the first investigations we will undertake and several new measures will be explored. A version of co-occurrence of concepts is one possibility where two concepts are said to co-occur if they appear as constituents within the same complex concept or if they appear in different complex concepts which share an edge in the DAG. Weighted co-occurrences where weights diminish by path-length between complex concepts will also be explored.

Information based semantic similarity measures have been used with some success in other contexts by Resnik (1999) but are not immediately applicable here. Unlike the taxonomic problems considered by Resnik (1999), upstream position in the DAG here does not correspond to increasing probability. Moreover there is as yet no database of empirical information available that would suit construction of relevant probability estimates (though this may change over time as Primal Fusion's user base grows). Nevertheless, we will look to attach a reasonable probability interpretation to the co-occurrence of constituent concepts in the DAG. We have already begun exploring how the probability calculus might be used in relating constituent concepts over the DAG by making certain plausible assumptions about conditional independence. This will be made more formal as part of this project with an eye towards incorporating empirical information as it becomes available.

Our approach will be to adapt at least these three different approaches to similarity to the problem as abstracted so far. This is non-trivial but should yield novel methods of more general applicability.

Given several competing similarity sources, we intend to explore two possible means of their combination. First, different similarities could be combined as a weighted sum, as for example in Rodriguez and Egenhofer (2003); choosing problem appropriate weights will be a particular challenge. Second, the combination could be made *after* clustering – *by combining the clustering results* from different similarity measures as opposed to combining the similarity measures. This is where the theory developed by Wu Zhou in his Ph.D. thesis will be directly applicable. Different similarity measures track different senses of similarity. Different clustering methods track different cluster structures. An advantage Zhou's approach is that both can be combined in the sense that 3 similarities x 2 clustering methods = 6 clusterings that can then be "averaged" to produce a single clustering. Because Zhou's method places clustering within a statistical framework, a quantitative measure of the variability across clustering outcomes will be possible. This opens up a number of possibilities for experimental investigation of ontologies.

A separate thread of this project, to be pursued primarily by graduate students and Waterloo under Prof. Oldford's supervision, will be to explore methods for how best to visualize and dynamically explore

the relations between concepts. Here the work of Hurley and Oldford (2010) will be adapted to this problem.

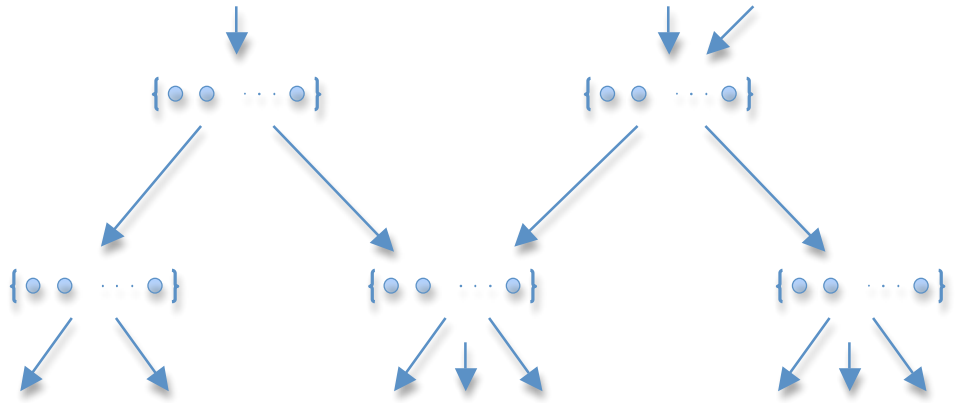
This project will bring statistical methods in general, and clustering methodology in particular, to bear on the semantic representations developed by Primal Fusion as part of their software platform. It will conduct scientific experiments within the domain of Primal Fusion's rich source of data. This is a rare opportunity to engage in research of this kind with access to this information. Research results here will apply to other numerous other domains and will appear in the general research literature.

Further details

The research outlined above has three separate thrusts. The first is to determine an appropriate measure of semantic similarity. The second is to determine how best to use statistical clustering methods in this context, taking advantage of the new theory developed by Zhou (2010). The third is to explore how to best visualize the results. These three can be considered in parallel or serially. The plan is to consider them serially while being mindful of the interaction between the topics and the effects this might have on achieving the best results.

Semantic similarity

The problem is to determine semantic similarity of one simple concept to another where simple concepts are organized in sets and each set of concepts appears as a node in a directed acyclic graph. The structure is sketched in the figure below:



Each filled circle is a "simple concept", each node a set of such concepts. Each simple concept can appear in more than one node. The goal is to derive a measure of similarity between any pair of simple concepts.

Intuitively, there are a number of properties that a measure of similarity between concepts might be expected to have. For example, concepts that appear in the same node are more similar to one another than are concepts that appear in different nodes. Concepts in different nodes that have a path between

them should have higher similarity than concepts in different nodes that are not connected by a path. That similarity should be greater the shorter the path between nodes. The greater the number of paths between concepts, the greater might be the similarity. For concepts in two separate nodes connected by an edge, the similarity between concepts should be stronger if each node contains fewer concepts than if it contains many concepts. These and other considerations can lead to a number of ad hoc definitions of similarity between concepts.

Our first task will be to understand the approach already taken by Primal Fusion and represent it in this framework. A second task will be to develop and to explore alternative similarity measures. Here two approaches suggest themselves.

One is to cast each concept immediately into a vector whose dimension is the total number of simple concepts and whose entries count the number of "co-occurrences" weighted according to whether the concepts occur in the same node or in different nodes (with weights diminishing the farther apart are the two nodes). Cluster analysis can then be applied to the cosine similarity matrix of these vectors.

Another is to directly determine similarities between concepts on a more formal basis. A novel approach that will be explored is to use a probabilistic representation based directly on the graph structure. For example, for concepts a and b appearing in the network we introduce the notion that concept a "leads to" b, written $a \rightsquigarrow b$, and consider how we might determine from the network alone what the probability is that $a \rightsquigarrow b$. Writing this probability as $p_{a \rightsquigarrow b}$, we hope to determine from the network reasonable estimates of $p_{a \rightsquigarrow b}$ and $p_{b \rightsquigarrow a}$ for all pairs of concepts a and b. The approach we will explore is sketched in Oldford (2010, attached) and exercises the fundamental rules of probability, e.g.

$$\begin{aligned} p_{a \rightsquigarrow b} &= \sum_i Pr(a \rightsquigarrow b, Condition_i) \\ &= \sum_i Pr(a \rightsquigarrow b \mid Condition_i) \times Pr(Condition_i) \end{aligned}$$

The conditions are defined by moving through the DAG. Repeated application of Bayes Theorem, the inclusion-exclusion rule, and a few simplifying assumptions allows some reasonable approximation of $p_{a \rightsquigarrow b}$ to be determined. (See Oldford, 2010, for details.) The main attraction of this approach is that it would give a clear theoretical foundation for similarity of concepts. The probabilities could be used directly as (asymmetric) similarity measures or used to provide information theoretic measures of similarity,

$$Sim(a, b) = -\log(p_{a \rightsquigarrow b})$$

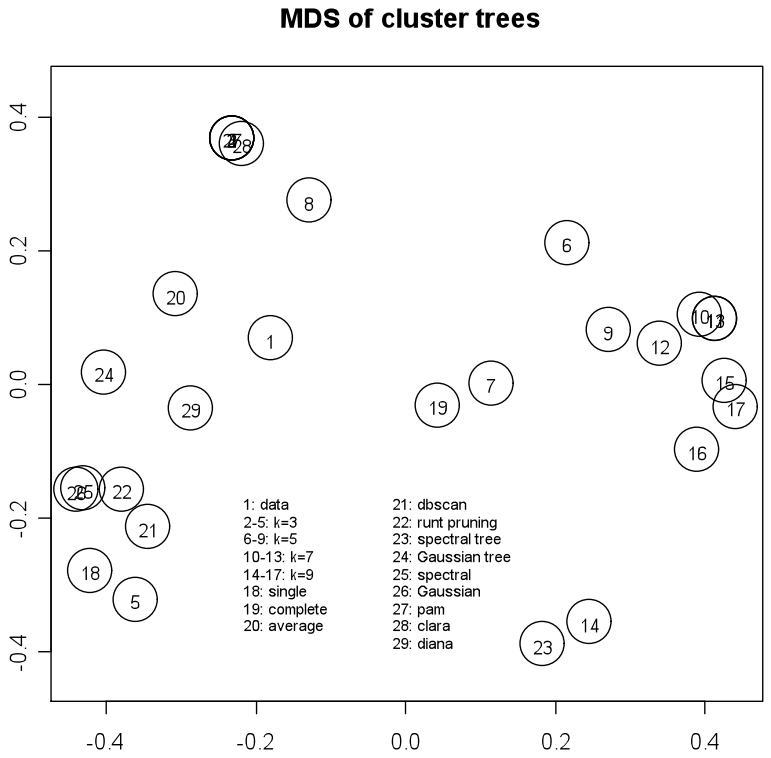
as in Resnik (1999) or as in Maguitman et al (2005). Again these could be used directly in clustering or first cast into a vector with the similarities as entries and then the cosine measure used.

It is clear that there will be several competing measures of semantic similarity. Combining these into a single measure (e.g. as in Jiang and Conrath, 1997, or Rodriguez and Egonhofer, 2003) will also be considered.

Cluster analysis

Given any measure of similarity, there are numerous clustering methods to choose from. Zhou's (2010) Ph.D. thesis provides a theoretical framework that allows the outcome of different clustering methods (including both hierarchical and partition methods) to be compared, to tell which clustering outcomes are similar and which are different. Zhou (2010) does this by providing a distance metric between finite cluster trees.

For example, the figure at right shows the results of 29 different clustering methods on a single data set (the "olive oil" data consisting of measurements on various fatty acids of different olive oils from nine separate olive growing regions in Italy). Each circle corresponds to the output of one clustering method on this data. Zhou (2010) defines a metric space for cluster trees that provides distances between clustering outcomes. The figure at right shows the clustering outcomes positioned in a two-dimensional space that roughly preserves these distances (here by multi-dimensional scaling, MDS). The "true" tree, that organizes the olive oils according to their known growing regions, is shown as the circle containing 1.



Near neighbours in this two-dimensional space are 20 and 29; these respectively refer to average linkage clustering and the divisive hierarchical clustering method called DIANA (e.g. see Kaufman and Rousseeuw, 1990). Knowing that 1 is the target, these methods seem to do best of the 28 different methods considered. Note that k-means clustering (14 through 17) with the "correct k" and different random starts does not perform at all well on this data set. Zhou's methodology allows us to compare one clustering to another and (in experimental conditions at least) to compare each clustering to the known, or target, clustering.

Moreover, the theory also allows for the combination of several distinct clustering methods into a single combined tree. This is particularly important to the present proposal. Not only can different clustering methods be selected that capture different group structure in the data, but the same clustering method used with different similarity measures can be used. This is an instance of interaction between the determination of similarity and the selection of clustering methods – many similarity methods and their

Data Visualization

Primal Fusion's application directs newly synthesized and inter-related concepts to the user to be further explored, examined, and assessed by the user. This requires a means of visualizing the various concepts and their inter-relations. A simple display is a tag-cloud where the tags are the words that define the concept. Alternatively, the tags might be images or any other representation of the concept.

Prof. Oldford and his students, in conjunction with Zhou and Primal Fusion, will also work on the development of new visualization tools to empower exploration of these synthesized concepts. The key ingredients here are relevance measures of each concept (vis-à-vis the user input) and the measures of similarity between concepts. Relevance can be used to determine the relative prominence of each tag and the similarity to determine relative positioning of each tag.

Mathematically, the positioning problem is that of finding a low-dimensional (e.g. two dimensions) positioning of the concept tags such that the inter-tag distances closely match the inverse of the similarity between the corresponding concepts. This is a dimensionality reduction problem for which there exist numerous methods now in the machine learning literature. The MDS procedure shown above for cluster methods is one such procedure; this and others will be explored.

One problem with tag-clouds is that restriction to two dimensions will necessarily result in loss of information about the inter-relations of the corresponding concepts. A more faithful representation would mean embedding the tags in a cloud of higher dimension, e.g. perhaps five dimensions. The difficulty with dimensions higher than two or three however is that they are not so easily visualized. Recent work by Hurley and Oldford (2010) however provide a means of navigating through higher dimensional spaces along natural 3-dimensional pathways. General software to navigate high-dimensional data in this way has recently been implemented by Oldford and his Ph.D. student Adrian Waddell. Further research on how this might be used to visualize tag-clouds in higher dimension will be explored as part of this project.

References

Hurley, C.B. and R.W. Oldford (2010). "Graphs as navigational infrastructure for high dimensional data spaces", to appear in *Computational Statistics*, 21 pages. Available at www.stats.uwaterloo.ca/~rwoldfor/papers/Accelerate/NavigationPaper.pdf

Jiang, J.J., and D.W. Conrath (1997). "Semantic Similarity Based on Corpus Statistics and Lexical Taxonomy", *Proc. Int'l Conf. n Comp. Linguistics (ROCLING X)*, Taiwan.

Kaufman, L. and P. Rousseeuw (1990). *Finding Groups in Data: An Introduction to Cluster Analysis*, Wiley-Interscience, New York.

Maguitman, A.G., F. Menczer, H. Roinestad and A. Vespignani (2005). "Algorithmic Detection of Semantic Similarity", *Proc of the WWW 2005*, Chiba, China. Pp. 107-116.

Oldford, R.W. (2010). "Towards a probabilistic model for semantic similarity on concept set DAGs", *Manuscript in preparation (pre-publication, www.stats.uwaterloo.ca/~rwoldfor/papers/Accelerate/graph.pdf)*, 7 pages.

Resnik, P. (1999). "Semantic Similarity in a Taxonomy: An Information-Based Measure and its Application to Problems of Ambiguity in Natural Language", *Journal of Artificial Intelligence Research*, *11*, pp. 95-130.

Rodriguez, M.A. and M.J. Egenhofer (2003). "Determining Semantic Similarity among Entity Classes from Different Ontologies", *IEEE Trans. On Knowledge and Data Engineering*, *15*, No. 2, pp. 442-456.

Tversky, A. (1977). "Features of Similarity", *Psychological Review*, *84*, No. 4, pp. 327-352.

Zhou, W. (2010). *A New Framework for Clustering*. University of Waterloo, Ph.D. dissertation, 225 pages.

3. Relevance to Primal Fusion

Primal Fusion will benefit from a knowledgeable critique of our techniques and methods, and from fresh insight into our key areas of semantic clustering and data management. Also, a working relationship with a post-doc could help Primal Fusion research staff members remain current. Semantic similarity is a core function of the Primal Fusion platform. The post-doc's expertise in clustering and statistical analysis will help us address multiple challenges that we currently face. The post-doc will produce research papers and code libraries that will benefit Primal Fusion in the longer term.

4. Expected interaction with Primal Fusion

Wu is expected to spend 50% of his time working on Primal Fusion project. He should primarily spend that time in our office. He will work offsite at the University of Waterloo with Prof. Oldford (Wu's Ph.D. supervisor) as well. Professor Oldford's research interests in computational statistics in general, and in clustering and data visualization in particular, is of clear immediate and long-term value to Primal Fusion. Other graduate students of Professor Oldford may visit and interact with Primal Fusion staff over the course of the project.

5. Research milestones and timeline

We plan to conduct this project as three four-month-long increments.

Months 1 through 4:

- Become familiar with and abstract the methods currently used in the analysis phase.

- In particular, “analysis” now implicitly constructs similarity measures of concepts in several ways including an ad hoc combination of feature matching and path distance. These need to be made more explicit and compared and contrasted with potential alternatives.
- More generally, assess soundness, provide a critique of each component, suggest existing alternative techniques that we could use for each part as well as the whole system.
- Document this work (for example, as a white paper), describing in general terms the present approach and possibly outlining new directions. The latter would be the more research intensive part.

This would get Wu onboard, give Primal Fusion a good whitepaper, and will spawn the next step which is to re-write the analysis engine and improve it.

Months 5 through 8:

- This will largely depend on progress made in months 1 through 4.
- It is anticipated that in this segment focus will be on the development of new similarity measures and the use of clustering methods in this context. To date Primal Fusion has not formally incorporated clustering methods, so experiments on what will be the best use of this technology will be critical.
- Investigations of how best to integrate the synthesis (user directed) and analysis phases will be begun.

Months 9 through 12:

- Again, this will largely depend on progress made in previous months.
- Supposing some significant success by month 8, a couple of research papers based on the work in months 1 thru 8 should be forthcoming to be submitted to appropriate research conferences.
- The next phase of research might be begun. This would likely be again looking at similarity measures and clustering methods appropriate as new ontologies are processed which different significantly in structure and information. Methods may differ depending on the ontology.
- Ultimately the problem of dynamic change will have to be addressed. This will be kept in mind throughout all research in all previous months, but will not likely be addressed head on until the earlier challenges are met.

6. Budget for Research Funds

We are applying for a double award so as to support a post-doctoral fellowship for a full 12 months. A full two thirds of the funds ($60,000) will be applied to salary (including all benefits) for the post-doctoral fellow, Wu Zhou. The remaining funds will be used to support other graduate students at the University of Waterloo under Prof. Oldford's supervision. These graduate students will primarily be working on topics in data visualization for this project. We allocate funds for 2 graduate students.

BUDGET	Months 1 to 4	Months 5 to 8	Months 9 to 12	Total/Year
SALARIES (including benefits)				
Post-doc	$20,000	$20,000	$20,000	$60,000
Grad. Student Research Assistants (Ph.D., Master's)	$9,000	$9,000	$9,000	$27,000
Subtotal	$29,000	$29,000	$29,000	$87,000
DIRECT COSTS				
Travel*	–	–	$2,500	$2,500
Computing**	$500	–	–	$500
Subtotal	$500	–	$2,500	$3,000
Total Cash Contribution from Primal Fusion				$45,000
Total Amount requested from MITACS				$45,000

Miscellanea

1. **Will the intern be applying for an additional travel subsidy over and above what may be in the proposed budget?** *If so, please outline projected expenses.*

   Yes___ No_x_

2. **Relationship to past/other MITACS or ACCELERATE internships where relevant**

   None

3. **Is the academic supervisor an owner or a co-owner of the partner organization, or does the supervisor participate in the day-to-day management of the organization?** *If so, please complete the Conflict of Interest Declaration.*

   Yes___ No__x _

   NO - however, the academic supervisor, Wayne Oldford, has stock options valued at 0.11% of the total company value.

4. **Will the proposed research be taking place outside of the lab or normal business environment?**

   Yes___ No_x__

5. **Does the proposed research involve living human subjects or human remains, cadavers, tissues, biological fluids, embryos or foetuses?** *If so, the proposal must be reviewed by the participating University Research Ethics Board, and a report by such board must be forwarded to MITACS as soon as available. (Please note no funds can be released until MITACS has received the report.)*

   Yes___ No_x__

6. **Does the proposed research involve animal subjects?** *If so, the proposal must be reviewed by the participating University Animal Care Committee and a report from such a committee must be forwarded to MITACS as soon as available. (Please note no funds can be released until MITACS has received the report.)*

   Yes___ No_x__

7. **Is a biohazards review required?** *If so the necessary review/report from your University must be forwarded to MITACS when available. (Please note no funds can be released until MITACS has received the report.)*

   Yes___ No_x_

8. **Are a) the industry funds of this internship being or going to be leveraged by another organization and b) will the proposed intern also be receiving salary in addition to the internship stipend through Tri-Council funds?** *If so, where applicable please provide a) the name of the organization and the amount being leveraged b) the Tri-Council agency, the Grant / Award name, and the annual salary or stipend which the intern will be receiving from the Tri-Council for the year including the internship period.*

   Yes___ No_x__

Suggested Reviewers:

1. **Prof. Hugh Chipman**
   Dept. of Mathematics and Statistics
   Acadia University
   12 University Avenue
   Huggins Science Hall
   Wolfville, Nova Scotia
   B4P 2R6
   Phone: 902-585-1525
   Fax: 902-585-1074
   Email: chipman@acadiau.ca

2. **Prof. Shirley Mills**
   School of Mathematics and Statistics
   Carleton University
   5203 Herzberg Building
   Ottawa, Ontario
   K1S 5B6
   Phone: 613 520 2199
   Fax: 613 825 6978
   Email: smills@math.carleton.ca

3. **Prof. Werner Stuetzle**
   Department of Statistics
   College of Arts & Sciences
   Box 353764
   University of Washington
   Seattle, WA 98195-3765
   USA
   Phone: 206 616 8709
   Fax: 206 543 5462
   Email: wxs@u.washington.edu

MITACS Accelerate Memorandum

The participants listed below have agreed to set in place an Internship based upon the attached proposal. It is understood that the sponsor organization contribution shall be provided to MITACS Inc. prior to commencement of the Internship, and that upon scientific approval MITACS shall forward the funds to the university as a research grant to the supervising professor, and that the internship stipend shall be paid to the student by the university from the grant. . MITACS is unable to assume liability for accidents, illness, or losses that may occur during the internship period. All parties are responsible for ensuring that they have appropriate insurance. All parties also agree that the intern is expected to provide MITACS with a completion report and all participants will complete an exit survey within 1 month of project completion.

All parties involved with MITACS Accelerate are bound by the standard intellectual property terms of the university where the intern is enrolled; except where Intellectual Property is covered by separate agreements to which the university and the Sponsor Organization are parties and which are active during the dates of the Internship. By signing this memorandum, you are acknowledging that you agree to the terms of the university where the intern is enrolled.

Please go to the following link:
http://www.mitacs.ca/index.php?option=com_content&view=article&id=246&Itemid=117&lang=en&limitstart=6 and click on "Intellectual Property", for any university specific IP policies regarding Accelerate internships.

The participants listed below also agree that MITACS can disclose personal information included in this proposal to the program's funding partners for the purpose of evaluating the Program and its outcomes.

Title of project: Clustering methodology for thought networks

Period of Internship: May 1, 2010 – April 30, 2011

Overview of Project: Primal Fusion works in the area of consumer-directed creation of semantic networks. A motivating conceptual model is to imagine posing the question "What are you thinking?" and (whatever the user's answer) to be able to deliver a rich semantic network of relevant concepts ranging from those the user will recognize as strongly related to those which might only be loosely related, or surprising, even thought provoking. This project will bring statistical methods in general, and clustering methodology and data visualization, in particular, to bear on the semantic representations developed by Primal Fusion as part of their software platform. Novel measures of semantic similarity will be developed and the recent Ph.D. research of the candidate will be applied in this industrial context. Novel methods of data visualization will also be applied to provide dynamic and interactive new interfaces for users to explore relations between concepts/topics of their interest.

Participants

Intern: Dr. Wu Zhou
Dept. of Statistics & Actuarial Science
University of Waterloo

Signature Date

Supervisor: Prof. R. Wayne Oldford
Centre for Computational Mathematics in Industry & Commerce
University of Waterloo

Signature Date

Organization

Sponsor: Dr. Ihab Ilyas
Position Vice-President of Research
Organization Primal Fusion Inc.

Signature Date

Office of Research Services

Representative: Kelly Moran
Position Manager, Grants and Government Research Contracts
Office of Research, University of Waterloo

Signature Date

MITACS Inc.
Representative: Claudia Krywiak
Position Vice-President, Business Development – Ontario

Signature Date