Linking Tacit Knowledge in the Pediatric Pain e-Mail Archives and Explicit Knowledge in PubMed

Zhixin Chen, Michael Shepherd and Syed Sibte Raza Abidi
Faculty of Computer Science
Dalhousie University
Halifax, NS, Canada B3H1W5
{zhixin,shepherd,sraza}@cs.dal.ca

G. Allen Finley
Departments of Anaesthesia & Psychology
Dalhousie University, Halifax, NS, Canada
allen.finley@dal.ca

Abstract

The goal of this research is to provide a means by which the tacit knowledge in an e-mail archive can be browsed in an organized manner and linked directly to the explicit knowledge found in PubMed. The Pediatric Pain Mailing List (PPML) is an international Internet-based forum for informal discussion of any topic related to pain in children. There are now over seven hundred members, including clinicians, researchers and patients from at least forty countries on six continents. Currently, the archive contains more than ten thousand messages. This paper reports on SOM-based approaches to the taci-knowledge organization and an MMTx-based approach that links the e-mail threads directly into the MeSH hierarchy and links to PubMed.

1. Introduction

The existence of pediatric pain has been recognized widely in the past fifteen years. Although knowledge about pediatric pain has accumulated over this period, it is still under-treated. Many pediatric pain problems are relatively rare and it is impossible for even a pediatric pain specialist to have personal experience with all possible symptoms and syndromes. In addition, clinical research has been limited and there is little in the medical and scientific literature to support clinical decision-making. Therefore, the Pediatric Pain Mailing List (PPML) has become an important valuable information resource for clinicians, researchers and patients.

The Pediatric Pain Mailing List (PPML) is an international Internet-based forum for informal discussion of any topic related to pain in children. It is a list server that permits individuals to post/reply to e-mail messages that are then sent to all the subscribers and the messages themselves are archived at the central server. It was developed ten years ago to promote informal discussions and information exchanges among pediatric professionals [5]. The subscribers to PPML include professionals from different clinical and research disciplines and laymen, such as medical students, pediatric patients and their parents. There are now over seven hundred members from at least forty countries on six continents. Currently, the archive contains more than ten thousand messages.

Over these ten years, a vast amount of tacit knowledge about pediatric pain has been captured in the PPML archives. Tacit knowledge is what the knower knows, knowledge that is derived from experience [11]. This tacit knowledge is captured and shared through e-mail discussions carried on amongst clinicians, researchers and patients through the PPML. As such, the PPML archives have become a pediatric pain knowledge repository.

Unfortunately, this knowledge is not readily accessible. The messages are stored in a raw e-mail format and are not organized in any manner. This has made it difficult to provide the subscribers with information retrieval and knowledge discovery from the archive. In order to provide access to this tacit knowledge, a thread-based method for clustering the messages into hierarchical categories was developed. The stored e-mail messages were first cleaned, then organized into discussion threads. These discussion threads were then clustered hierarchically using repeated applications of the *k*-means algorithm [16]. The resulting hierarchy was evaluated against categorizations by two human experts and there were no significant differences found between the results of these two evaluations [14].

0-7695-2507-5/06/$20.00 (C) 2006 IEEE

The overall goal of this research is to link the tacit knowledge contained in the PPML to the explicit knowledge contained in a dataset such as PubMed. A previous approach, based on term extraction and *k*-means clustering, is shown in Figure 1. While the results of the previous research [14] provided a hierarchical clustering of the message threads and a web-based interface for browsing the clusters, it did not map the concepts from the PPML into MeSH and did not link the threads to PubMed.

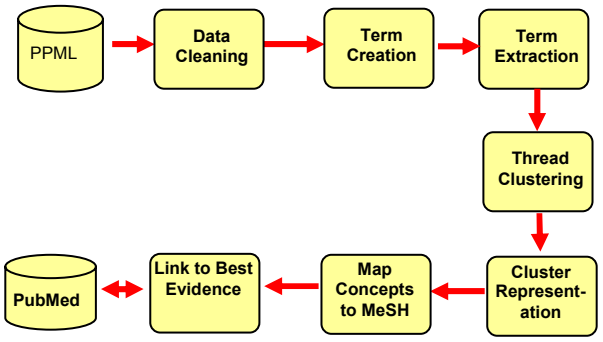

Figure 1. Steps to link PPML and PubMed

Although the previous results allow the user to browse the cluster hierarchy and to select different hierarchical organizations (broad but shallow vs. narrow but deep), it produces a set of terms representatives for each cluster that does not map easily to MeSH. Therefore, in the current research, *k*-means clustering was replaced with a Self-Organizing Map (SOM) for clustering the threads on the basis that SOM might improve the quality of the clustering and be better able to identify those terms that were most responsible for cluster creation as good cluster representatives. The resulting cluster organizations were then evaluated but were found to be not as good as the *k*-means clustering.

Therefore, we changed our methodology (Figure 2) so that terms were extracted from the threads using MMTx [9] and mapped directly into the MeSH hierarchy. The threads from which the terms were extracted were then associated with those places in the MeSH hierarchies. In other words, the threads were not clustered, rather they were classified according to the MeSH hierarchies. Based on these MeSH mappings, the individual threads were linked into PubMed.

Section 2 of this paper discusses the PPML within the context of tacit and explicit knowledge and describes the previous research [14]. Section 3 describes the SOM methodology and resulting evaluations. Section 4 describes the MMTx extraction of terms and mapping into MeSH. Section 5 summarizes this paper and discusses future work on linking the PPML to the explicit knowledge found in the published literature.

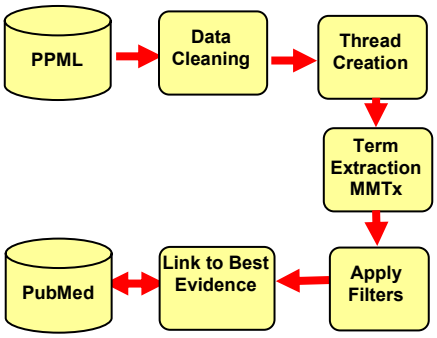

Figure 2. Steps using MMTx

2. Tacit-Explicit knowledge framework

Polanyi [10] introduced the notions of tacit knowledge and explicit knowledge and Nonaka [11] used these notions to formulate a theory of organizational learning. Polanyi's and Nonaka's work can be used to frame the research of the PPML research project, i.e., to frame the knowledge held in the PPML archives and the methods being developed to access this knowledge.

Tacit knowledge is what the knower knows and is derived from experience whereas explicit knowledge is represented by some artifact such as a document or journal article [8]. Within the PPML, the tacit knowledge of clinicians, researchers and patients with respect to pediatric pain is captured and transformed into explicit knowledge as represented by threads of e-mail messages that represent conversations and discussions on particular topics. This transformation of knowledge is represented by Figure 3, as adapted from Nonaka. The four processes by which knowledge is transformed are socialization, externalization, combination and internalization. The spiral indicates the continuous nature of this transformation among the four processes. The processes are:

- Socialization is the transformation of tacit knowledge to tacit knowledge as experienced in face-to-face meetings and the communication is **synchronous** in nature;
- Externalization is the transformation of tacit knowledge to explicit knowledge as might take place in responding to questions. In this instance, the communication may be **asynchronous** in nature as found in list server discussion groups;
- Combination is the transformation of explicit knowledge to explicit knowledge. Once knowledge as been articulated and captured in a persistent form, it can be organized into a hierarchically organized set of categories where the organization itself reflects knowledge of the domain;

- Internalization is the transformation of explicit knowledge to tacit knowledge. Access to an organized or structured set of documents that is a representation of explicit knowledge helps the user to more quickly internalize the knowledge, thus converting the explicit knowledge into their own tacit knowledge.

Within the PPML, there is no accommodation for the process of socialization, i.e., the transformation of tacit knowledge to tacit knowledge as there is no means for **synchronous** communication as might be found in chat rooms or supported by groupware. The PPML does, however, support the process of externalization as tacit knowledge is elicited as people ask questions through the list server and those with tacit knowledge respond, **asynchronously**. The tacit knowledge captured is that of the responder, i.e., what the responder knows from experience. The post-reply dialog transforms this tacit knowledge into explicit knowledge. These e-mail discussions or threads are artifacts of explicit knowledge.

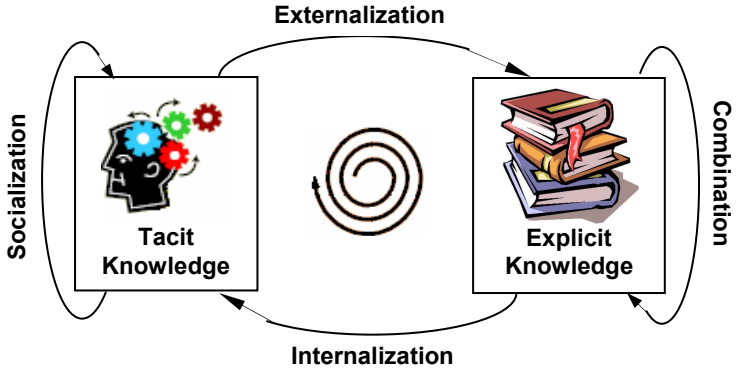

Figure 3. Knowledge transformation processes

3. Previous PPML research

Unfortunately, the knowledge accumulated in the PPML archives was very difficult to access as it was not organized in any manner other than chronologically. Therefore, the previous research [14] transformed this explicit knowledge by cleaning the data and tying together e-mail posts and replies into e-mail threads. Prior to cleaning and threading, there were 6939 messages in the archive, from 1993 to 1999. After cleaning and threading, this was reduced to 4033 messages in 1289 threads. A sample thread is shown in Table 3 with all identifying information removed.

These threads were then clustered hierarchically using the *k*-means clustering algorithm [4]. This hierarchical organization is the "valued-added" by this explicit to explicit knowledge transformation (shown in Figure 4). This hierarchically organized structure permits users to browse through the knowledge captured in the PPML archives and to internalize this explicit knowledge (internalization).

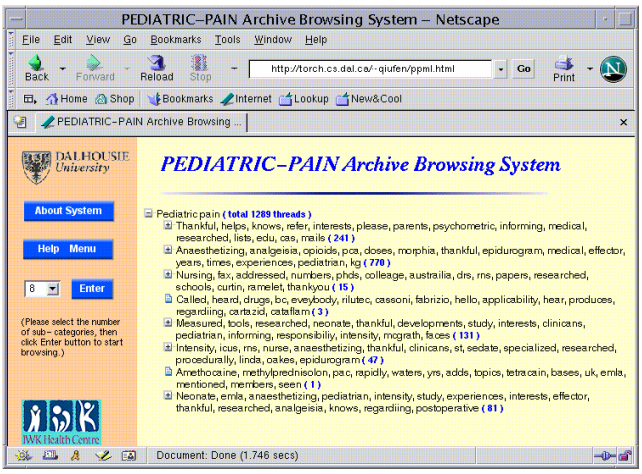

Figure 4. The hierarchical organization of the Pediatric Pain archive

One hundred threads were selected randomly and clustered using the term selection and clustering algorithms developed in that research and by two subject experts, independently. Expert 1 manually created 13 clusters and expert 2, 17 clusters. An information theoretic measure of redundancy [18] showed that there was some similarity between the clusters generated by the two experts. The algorithm-based clustering was evaluated against each of the experts using the *F*-measure and the results showed that there was no significant difference between the two sets of manually generated clusters on the overall *F*-measures. The average overall *F*-value measured relative to expert 1 was 0.47 and relative to expert 2 was 0.48. From this, one can infer that computer-generated clustering is not that different from either of the two manually generated sets of clusters, and is a reasonably good representation of the organization of the domain of explicit knowledge represented in the threaded dataset.

4. SOM Methodology and Results

The previous research [12], summarized in Section 3, generated a large number of term representatives for each cluster with no elegant way of mapping the terms into MeSH. Therefore, the *k*-means clustering algorithm was replaced with a SOM [7] in the expectation that the clustering results would be better and that a smaller set of term representatives for each cluster might be identified.

The threads were represented by term vectors of size 4111. Principal Component Analysis (PCA) was performed to map this set of vectors into a smaller vector space. This reduced the number of features significantly and experiments were performed with different numbers of features and different numbers of cells in the SOM.

The results of the PCA can be seen in Figure 5. The x-axis represents the eigenvectors and the y-axis the eigenvalues. The eigenvalues dropped off quickly and approached zero and stabilized at the 150th eignvector, indicating that almost all of the information can be captured by the first 150 eigenvectors or features. This was verified by the experiments described below in which a wide range of the number of PCA features were used and the best results occurred for the first 150 features.

Three different approaches to using the SOM for clustering were investigated, and the results can be viewed in Table 1. In all instances, the same 100 threads were used as for the previous research [12] and the same *F*-measure was used to evaluate the results relative to the two sets of clusters generated by the two domain experts. For comparison purposes, Table 1 also contains the average *F*-measures obtained from the hierarchical *k*-means clustering from the previous research, with *k*=6.

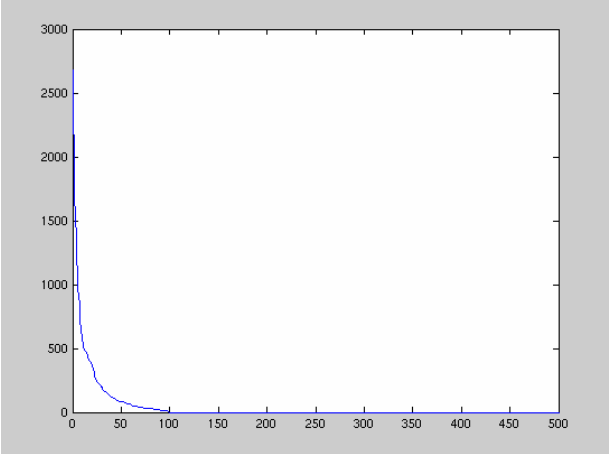

Figure 5. PCA reduction of feature set.

4.1 First SOM approach

The first approach treated each individual cell in the resulting SOM as an individual cluster. A number of experiments were run, varying the number of features from 100 to 4111 (all features), the number of cells in the SOM and the dimensions of the SOM from as small as a 2*2 to a 8*6 map.

As can be from Table 1, the best *F*-measures when evaluated relative to the two experts for this approach were for an 8*6 map and 150 features, but the results were not as good as the average results for the *k*-means approach from the previous research.

4.2 Second SOM approach

The second approach created a hierarchical SOM using the Growing Hierarchical SOM algorithm [4], again treating each resulting cell as a separate cluster. A number of experiments were run, varying the number of features, the number of layers or depth of the hierarchy, the number of sub-trees or branches at each level.

As can be seen from Table 1, the best results were for 500 features and a 5-layer hierarchy with 2*2 branching and a total of 53 cells populated. Although the results were better than for the first SOM approach, the results were only approximately equivalent (but not quite) as the average *F*-measure results for *k*-means approach from the previous research.

4.3 Third SOM approach

The third approach built on the first SOM approach by treating each cell as a unit and clustering these units into a fewer number but larger clusters. Each cell was represented by the centroid of the threads assigned to that cell and then the *k*-means clustering algorithm was applied to these units. This version of the *k*-means algorithm created a single layer of clusters, not a hierarchical clustering.

The number of features was fixed at 150 but the size and dimensionality of the SOM maps were varied. In all, 31 different maps were created. Once each SOM was created, the *k*-means algorithm was applied to that map. The value of *k* was varied from 1 through 20 and, for each value multiple iterations were run with *k* randomly selected units as seeds for the clusters. The Davies-Bouldin Index [15] was used to determine the highest quality clustering for each value of *k*.

The Davies-Bouldin Index is based on the density of each cluster and distance between clusters on the assumption that a good cluster structure would have clusters in which the elements of each cluster are close together and in which the clusters themselves are well separated. The *F*-measure was then calculated for each cluster identified by the Davies-Bouldin Index.

In Table 1, the results of this approach are identified as SOM-*k*. The best cluster structure was determined to be for a 10*5 SOM with a total of 13 clusters. Surprisingly, the *F*-measures generated were the lowest of all the approaches, even though the Davies-Bouldin index was applied and the resulting number of clusters was similar to the number of clusters created by the domain experts (13 and 17, respectively).

In summary, none of the above SOM-based approaches provided *F*-measures as good as the average *k*-means clustering from the original research.

Table 1. Comparison of best SOM-based results with *k*-means, *k*=6

Method	Features	Map Size	Number of Clusters	Best *F*-Measure (SOM)		Average *F*-Measure (*k*-means)	
				Expert 1	Expert 2	Expert 1	Expert 2
SOM	150	8*6	48	0.2968	0.4043	0.47	0.48
GHSOM	500	5 layers, 2*2	53	0.4235	0.4466		
SOM-*k*	150	10*5	13	0.2783	0.3896		

5. MMTx Based Term Identification and Filtering: Methodology and Results

As the results of the SOM-based approach were not as good as the *k*-means approach and we were not satisfied with the *k*-means results, it was decided to investigate a different approach. This approach was to use the MMTx tool [9] to identify term phrases that would map into the MeSH hierarchy.

MMTx parses text into components including sentences, paragraphs, phrases, lexical elements and tokens. It produces a shallow syntactic analysis with part-of-speech tagging. Variants are generated from the resulting phrases, including acronyms, abbreviations and synonyms. Candidate concepts from the UMLS Metathesaurus are retrieved and evaluated against the phrases. The best of these candidates are organized into a final mapping in such a way as to best cover the text.

The final result from MMTx includes lists of term phrases and their semantic types. For instance, a partial list of the results for the term "discharged" is:

```
Phrase: "discharged"
Meta Candidates (3)
  966 C0030685:Discharge <1> (Patient Discharge)
      [Health Care Activity]
  966 C0600083:Discharge <3> (Discharge, Body
          Substance, Sample)
          [Body Substance]
  966 C0012621:Discharge, NOS (Discharge, Body
          Substance)
          [Body Substance]
```

In this example, three candidates are returned. They each have the same score of 966 out of 1000 indicating that MMTx feels they are all good matches. The MeSH terms that were matched are indicated by the concept numbers, such as C0030685, and by the concept terms in round braces. The terms in the square braces are the UMLS Semantic Types.

5.1 Filtering the MMTx results

The main problem with the MMTx output is that for each phrase that it identifies as a medical term it returns a list of potential candidate MeSH terms. Hence, for each thread, MMTx returns hundreds of candidate terms, if not thousands. Of course, such a large number of terms renders any subsequent processing infeasible and also leads to the introduction of diverse interpretations of the original medical concepts within a thread. Our approach, therefore, was to design a term filtering mechanism to identify a small set of MeSH terms representative of a PPML thread. We designed three different term filters and the user has the functionality a particular filter.

Our term filtering approach operates at the semantic/conceptual level as opposed to the term level [6]. The UMLS semantic types associated with each MeSH term is used as the basis for term filtering. The rationale is that working at a higher level of abstraction—i.e. the semantic level—we can (a) establish a medical context for the thread which can assist in subsequent search for corresponding literature; (b) characterize the entirety of medical terms into a small number of medical concepts—134 semantic types to be exact [13]; and (c) design filtering rules that apply to broad semantic types as opposed to focused individual terms [6]. We have developed the following term filters that can be chosen by the user:

Filter 1: Filtering Based on Semantic Group: For this filter we make use of the pre-defined semantic grouping of the UMLS semantic types—the 134 semantic types are distinguished into 15 semantic groups [6] as shown in Table 2.

Our approach is to filter out terms that belong to semantic groups (the term to semantic group relationship is via the semantic types) that have nominal significance towards pediatric pain terms/concepts. Table 2 shows the list of semantic groups, the corresponding semantic types for each semantic group and whether terms belonging to the semantic group will be filtered or not.

In essence, we work with only three semantic groups—i.e. DISO, CHEM, ANAT. These groups are deemed relevant because: DISO contains the semantic types of concepts related to medical disorders; CHEM contains the semantic types of concepts related to chemicals & drug. These semantic types are quite pertinent to drug related discussions noted in PPML; and ANAT contains semantic types of concepts related to anatomy, thereby indicating the physical location of the problem.

Table 2. List of semantic groups

Semantic Groups		Semantic Types	Terms Retained
Activities & Behaviors	ACTI	Activity, Behavior, Event, Machine Activity …	NO
Anatomy	ANAT	Anatomical structure, Body location …	YES
Chemicals & Drugs	CHEM	Amino Acid, Antibiotic, Chemical …	YES
Concepts & Ideas	CONC	Classification, Concept Entity …	NO
Devices	DEVI	Medical Device, Research Device …	NO
Disorders	DISO	Acquired Abnormality, Disease …	YES
Genes & Molecular Sequence	GENE	Amino Acid Sequence, Gene or Genome, Molecular Sequence …	NO
Geographic Areas	GEOG	Geographic Area	NO
Living Beings	LIVB	Age group, Alga, Animal …	NO *
Objects	OBJC	Entity, Food, Manufactured Object …	NO
Occupations	OCCU	Biomedical Occupation …	NO
Organization	ORGA	Organization, Professional Society …	NO
Phenomena	PHEN	Biologic Function, Test Result …	NO **
Physiology	PHYS	Cell Function, Clinical Attribute …	NO
Procedures	PROC	Diagnostic procedure …	NO ***

* Except Age Group (T100)

** Except Laboratory or Test Result (T034)

*** Except Diagnostic Procedures (T060), Therapeutic or Preventive Procedure(T061)

Filter 2: Filtering based on MeSH Tree: This filter makes use of the MeSH trees to classify medical concepts. There are 15 different MeSH trees, each tree pertaining to a specific medical aspect, such as disease, drugs, etc. Again, we select the MeSH trees that are useful for classifying the PPDL discussions, and MeSH terms belonging to the selected trees are retained whereas the rest of the terms will be filtered. For our purposes, we work with C (Diseases) and D (Chemical and Drugs) trees. The rationale for selecting these two trees is as follows: (a) The C tree contains disease related MeSH terminology. Since PPML largely deals with disorders and diseases, it was determined to keep the C tree to retain

disease related terminology; and (b) The D tree contains MeSH terminology about chemicals and drugs. The PPML contains discussions on drugs hence it was deemed important to retain drug-related terminology.

Filter 3: Filtering based on Mapping Score of Title: This filter makes use of MMTx mapping scores as the basis for filtering out terms. Each phrase encountered by MMTx is transformed to a MeSH term with a mapping score that reflects the goodness of the mapping of the original phrase to a MeSH term. The mapping score ranges between 0 – 1000, where the maximum mapping score of 1000 indicates a perfect mapping. We believe that MeSH terms with a high mapping score are good candidates for subsequent search for literature and hence should be retained. However, we believe that it is also important to consider the physical location of the term; terms that appear in the title are more representative of the context of the discussion as compared to terms in the body of the message. We have designed the following filtering rules that are applied in the sequence that they are presented:

- For a term in the title, if mapping score = 1000 then retain the term.
- For a term in the title, if semantic type = Age group (T100) then retain the term.
- For a term in the title, if semantic group = CHEM | DISO | ANAT AND (mapping score > 800) then retain the term.
- For terms in the title, if semantic type = Diagnostic Function (T060) | Therapeutic or Preventive Procedure (T060 | Laboratory or Test Result (T034) AND (mapping score > 800) then retain the term.

Filter 4: Filtering based on Mapping Score of Body: This filter works similar to that of filter 3, with the difference being that the filtering rule is applied to the body of the messages. The filtering rules in order of precedence are:

- For concepts in the message body, if semantic group = CHEM | DICO AND (mapping score = 1000) then retain the term.
- For concepts in the message body, if semantic group = DICO AND (mapping score > 600) then retain the term.

5.2 Term Filtering: Working Example

In this section we provide a working example to illustrate the functionality of the abovementioned methods to link discussion threads (i.e. tacit knowledge) to medical literature at PubMed (i.e. explicit knowledge).

The threads are organized based on (a) the UMLS semantic network—shown in Figure 6 and (b) the MeSH tree—shown in Figure 7 [2]. Users can traverse through the hierarchical structure (of both representations) to select a thread of interest.

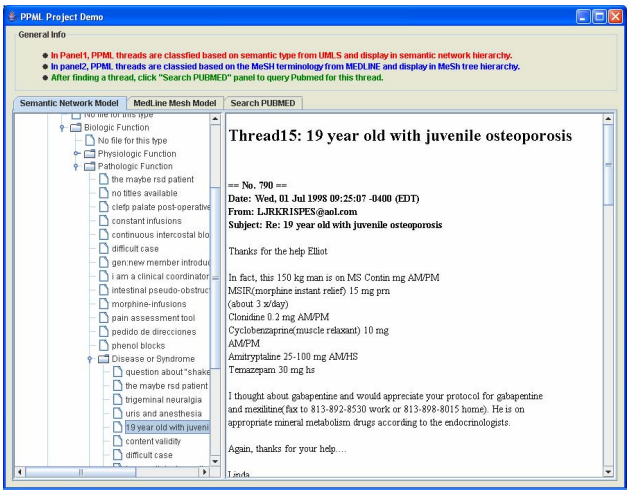

Figure 6. User interface showing the organization of the threads based on UMLS semantic network

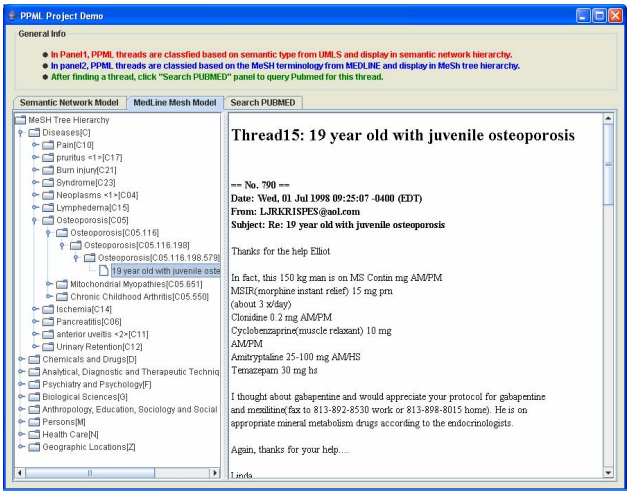

Figure 7. User interface showing the organization of the threads based on MeSH tree

For the purposes of this discussion, we select a discussion thread—*thread 15* that encapsulates a discussion pertaining to *19 year old with juvenile osteoporosis*—and work it through the various stages. Thread 15 contains 5 email messages exchanged between 3 pediatric pain practitioners (shown in Table 3).

Table 3. Content of thread 15

<><><><> Thread 15 <><><><> == No. 790 == Subject: Re: 19 year old with juvenile osteoporosis In fact, this 150 kg man is on MS Contin mg AM/PM MSIR(morphine instant relief) 15 mg prn (about 3 x/day) Clonidine 0.2 mg AM/PM

Cyclobenzaprine(muscle relaxant) 10 mg AM/PM
Amitryptaline 25-100 mg AM/HS
Temazepam 30 mg hs

I thought about gabapentine and would appreciate your protocol for gabapentine and mexilitine(fax to 813-892-8530 work or 813-898-8015 home). He is on appropriate mineral metabolism drugs according to the endocrinologists.

== No. 791 ==
Subject: Re: 19 year old with juvenile osteoporosis

If payment of an epidural catheter is an issue, then I suppose that a CAD-PCA pump would also not be paid for. Otherwise, a CAD-PCA pump infusion of an opioid such as hydromorphone or morphine is a possibility. Why not try Duragesic or an SR opioid such as OxyContin, titrated for pain relief. The paresthesias suggest central pain; thus, gabapentin (Neurontin) starting at 100 mg hs and titrated as necessary might also help. I'm not a consultant in pediatric pain per se, but I don't think the pediatric/adult split is an issue here.

== No. 792 ==
Subject: Re: 19 year old with juvenile osteoporosis

There is no need to start both gabapentin and mexletine at the same time. Try them sequentially. Also, the dosage of Morphine is quite low - I'd try titrating it as far as you can, until side effects become an issue, in whichyou can try other opioids.

== No. 921 ==
Subject: 19 year old with juvenile osteoporosis

I am seeing a patient tomorrow with this diagnosis who has been extensively cared for at the Medical University of South Carolina. He weighs about 120 Kg. He has had multiple fractures and episodes of severe back pain. In the past, he has been treated with Demerol, Doxepin, Flexeril for pain. He has also had clonidine and MS-Contin in the past. An epidural catheter was placed a couple of years ago with good pain relief, but insurance denied placement of an implanted catheter.

He now has a mreo sever episode of back pain, with paresthesias in both legs. Insurance will no longer allow him to go to the Med Univ of South Carolina, where the world expert endocrinologist on this disorder is. Of course, I am calling the pain folks there today, but wonder if anyone else has any experience with this disease?

== No. 922 ==
Subject: Re: 19 year old with juvenile osteoporosis

I know you've already thought of this, but it's worth stating that there is probably a role for TCAs in his management. In addition to TCAs and conventional analgesics and muscle relaxants, I would also add meds aimed at the alleviation of the neuropathic component of his pain, either gabapentin or mexilitine (need a protocol? email me.).
I assume he's already on endocrine Rx to improve his osteoblast activity.

The data pre-processing stage involved:

- Removal of the "Date", "From" and "Subject" tags from the message header.
- Removal duplicate content in the message body (happens when replying to a message)
- Removal of special characters, e.g. -, ~, ^, +, #, $
- Removal of the sender's contact information

In the term identification stage, the thread title and content is provided to MMTx to identify the underlying UMLS concepts. In total, 151 unique UMLS concepts are identified (a sample is shown in Table 4).

Table 4. A sample of MMTx output. Showing the semantic group, mapping score, concept and semantic type.

DISO|C0030554|1000|Paresthesia|T184|Sign or Symptom
DISO|C0234243|1000|Central pain|T184|Sign or Symptom
DISO|C0001688|1000|adverse effects|T037|Injury or Poisoning|T046|Pathologic Function
DISO|C0238740|1000|BACK PAIN, SEVERE|T184|Sign or Symptom
DISO|C0012634|1000|Disease|T047|Disease or Syndrome
CHEM|C0242402|1000|Opioids|T121|Pharmacologic Substance|T131|Hazardous or Poisonous Substance
CHEM|C0012306|1000|Hydromorphone|T109|Organic Chemical|T121|Pharmacologic Substance
CHEM|C0722364|1000|OxyContin|T109|Organic Chemical|T121|Pharmacologic Substance
CHEM|C0060926|1000|gabapentin|T109|Organic Chemical|T121|Pharmacologic Substance
CHEM|C0678176|1000|Neurontin|T109|Organic Chemical|T121|Pharmacologic Substance
CHEM|C0935966|1000|Demerol|T109|Organic Chemical|T121|Pharmacologic Substance
CHEM|C0013085|1000|Doxepin <1>|T109|Organic Chemical|T121|Pharmacologic Substance
CHEM|C0728797|1000|Flexeril|T109|Organic Chemical|T121|Pharmacologic Substance
ANAT|C1140621|1000|Leg <1>|T029|Body Location or Region

The rather large number of UMLS concepts identified demands the application of a term filter to filter out 'less relevant' concepts. We apply *filter 3* (as described earlier) to the title of the thread to identify the candidate terms. Table 5 shows the output of filter 3, note that five UMLS concepts are identified and subsequent filtering based on their semantic group results in only three concepts to be retained—i.e. *Feline osteogenesis imperfecta, Adolescent and Osteoporosis*.

Table 5. The filtering process via filter 3

Concept Name	Score	Semantic Group	Semantic Type	Filter
Year	694	CONC	Temporal Concept	Yes
Old	861	CONC	Temporal Concept	Yes
Feline osteogenesis imperfecta	1000	DISO	Disease or Syndrome	No
Adolescent	694	LIVB	Age Group	No
Osteoporosis	861	DISO	Disease or Syndrome	No

In the final stage, the set of UMLS concepts provided by the term filter are used to compose a search query for PubMed. In this case the search terms are {Feline osteogenesis imperfecta, Adolescent, Osteoporosis}.

A PubMed search can be further constrained by applying built-in features at PubMed. For instance, the user can (a) search by Clinical Study Category by specifying the query context in terms of therapy, diagnosis, etiology and prognosis [3,17]; and (b) search systematic reviews whereby the search attempts to find citations for systematic reviews, meta-analyses, reviews of clinical trials, evidence-based medicine, consensus development conference and guidelines [3,17]. Figure 8 shows our user interface for searching PubMed based on the selected thread (the query terms are automatically derived by the methods discussed earlier). Figure 9 is the result of a sample search.

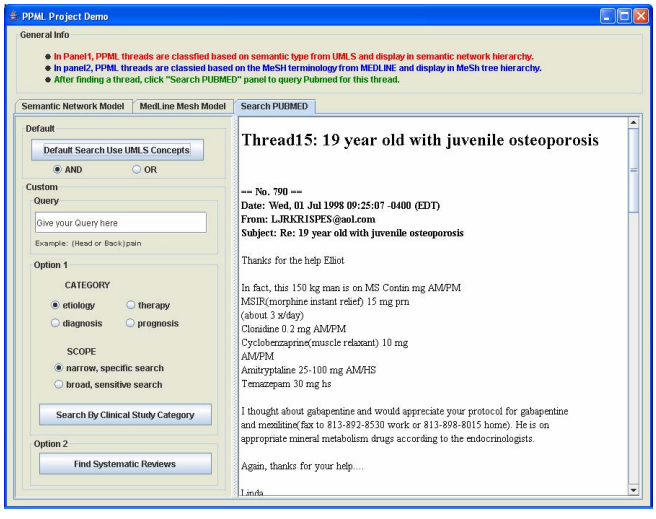

Figure 8. User interface for submitting thread-based search query to PubMed

Figure 9. A sample PubMed search query result

6. Summary and future work

The intent of the overall research project is to link the tacit knowledge captured in the PPML archives to the explicit knowledge as represented by best-evidence publications found in refereed journal and conference papers in service providers such as PUBMED [10]. As people browse the PPML, they may wish to discover best-evidence publications to either support or refute knowledge expressed in the PPML. The project will extract the main concepts from the thread or cluster of threads and use these concepts to search PUBMED automatically, returning any documents found.

Similarly, publications are often somewhat "dated". Therefore, people reading documents from PUBMED should be able to link directly to the PPML to discover if there are any recent e-mail discussions concerning a particular topic. The main concepts from the PUBMED article will be used to search for relevant threads in the PPML.

Previously [14], we demonstrated that the tacit knowledge of experts can be captured and transformed into explicit knowledge through asynchronous discussions and question-answering as provided by the PPML. This explicit knowledge is captured as e-mail threads. This explicit knowledge, represented as e-mail threads, can be further transformed through being organized hierarchically to represent the domain knowledge captured in the archives and expressed in the threads and that this hierarchical organization is a reasonably good representation. This explicit domain knowledge can then be browsed or queried by clinician, researchers and patients.

The research reported in this paper has taken a different approach to the problem in that it extracts concepts from the threads using MMTx and applies various filters to

reduce the number of extraneous concepts, maps the threads into MeSH hierarchies, and generates PubMed queries automatically.

There is still a lot of research remaining to be done, including full user evaluations to determine the efficacy of this method of discovering and linking tacit knowledge and explicit, best-evidence knowledge. The research to-date has focused on a one-direction mapping, from the PPML to PubMed. The research for mapping from PubMed to the PPML remains to be done and should prove equally as interesting as the research accomplished so far.

7. Acknowledgements

We would like to thank Dr. Carl von Baeyer, Department of Psychology, University of Saskatchewan, for manually creating one of the sets of clusters for evaluation purposes. The second set was created by Dr. Allen Finley, Departments of Anaesthesia and Psychology, Dalhousie University, one of the authors of this paper.

8. References

[1] Abidi S.S.R, Kershaw M. and E. Milios "BiRD: A strategy to autonomously supplement clinical practice guidelines with related clinical studies", *38th IEEE Hawaii International Conference on System Sciences (HICSS-38),* Hawaii, January 3-6 2005.

[2] Bodenreider, O. and A.T. McCray. "Exploring semantic groups through visual approaches", *Journal of Biomedical Informatics*, 36(6), 2003, pp. 414-432

[3] Clinical Queries (PubMed), NLM, USA, [http://www.ncbi.nlm.nih.gov/PubMed/clinical.htm] Available June 15, 2005.

[4] Dittenbach, M., Rauber, A. and D. Merkl Uncovering Hierarchical Structure in Data Using the Growing Hierarchical Self-Organizing Map. *Neurocomputing*, Vol. 48, No. 1-4, pp. 199-216, October 2002, Elsevier.

[5] Finley, G.A. and P.J. McGrath, "The PEDIATRIC-PAIN Mailing List – An Electronic Interdisciplinary Team", *3rd International Symposium on Pediatric Pain,* 1994.

[6] Haynes B, Wilczynski, McKibbon A, Walker C and J. Sinclair. "Developing optimal search strategies for detecting clinically sound studies in MEDLINE", *Journal American Medical Informatics Association*, 1994, pp. 447-458.

[7] Kohonen, T. *"Self-organized formation of topologically correct feature maps*," Biological Cybernetics, vol. 43, pp. 59-69, 1982.

[8] Marwick, A.D. "Knowledge Management Technology", *IBM Systems Journal*, 2001, 40(4).

[9] MetaMap Transfer Home [http://mmtx.nlm.nih.gov/], Available June 15, 2005.

[10] National Library of Medicine. [http://www.pubmedcentral.nih.gov/] Available June 15, 2005.

[11] Nonaka, I. and H. Takeuchi, *The Knowledge Creating Company*. Oxford University Press, Oxford, 1995.

[12] Polanyi, M., "The Tacit Dimension", In Prusak, L. (Ed.), *Knowledge in Organizations,* Butterworth-Heinemann, Woburn, MA, 1997, 135-146.

[13] Pratt. W. and L. Fagan. "The usefulness of dynamically categorizing search results", *Journal of the American Medical Informatics Association,* 2000, pp. 605-617.

[14] Qi, Q., Gao, Q., Shepherd, M. and G.A. Finley, "Accessing Tacit Knowledge in the Pediatric Pain e-Mail Archives", 38th Hawaii International Conference on System, Sciences, Hawaii, 2005.

[15] Stein, B., zu Eissen, SM. And Frank WiBbrock. "On Cluster Validity and the Information Need of Users", 3rd IASTED International Conferencer on Atificial Intelligence and Applications", Benamaldena, Spain, Septepber 2003.

[16] Steinbach, M., Karypis, G. and V. Kumar, "A Comparison of Document Clustering Techniques", *Proceeding of Text Mining Workshop, KDD,* 2000.

[17] SUMSearch, Society of General Internal Medicine, University of Texas, USA, http://sumsearch.uthscsa.edu/

[18] Tague, J. and M. Shepherd, "Redundancy as a measure of Classification Similarity", *Journal of Informatics*, 2(3), 1978, pp. 123-134.