Ontology-centric integration and navigation of the dengue literature

Menaka Rajapakse a,1, Rajaraman Kanagasabai a,1, Wee Tiong Ang a, Anitha Veeramani a, Mark J. Schreiber b, Christopher J.O. Baker a,*

a Institute for Infocomm Research (I2R), Data Mining Department, 21, Heng Mui Keng Terrace, Singapore 119613, Singapore
b Novartis Institute for Tropical Diseases Pte Ltd. (NITD), #05-01 Chromos, 10 Biopolis Road, Singapore 138670, Singapore

Abstract

Uninhibited access to the unstructured information distributed across the web and in scientific literature databases continues to be beyond the reach of scientists and health professionals. To address this challenge we have developed a literature driven, ontology-centric navigation infrastructure consisting of a content acquisition engine, a domain-specific ontology (in OWL-DL) and an ontology instantiation pipeline delivering sentences derived by domain-specific text mining. A visual query tool for reasoning over A-box instances in the populated ontology is presented and used to build conceptual queries that can be instantiated to the knowledgebase. We have deployed this generic infrastructure to facilitate data integration and knowledge sharing in the domain of dengue, which is one of the most prevalent viral diseases that continue to infect millions of people in the tropical and subtropical regions annually. Using our unique methodology we illustrate simplified search and discovery on dengue information derived from distributed resources and aggregated according to dengue ontology. Furthermore we apply data mining to the instantiated ontology to elucidate trends in the mentions of dengue serotypes in scientific abstracts since 1974.

© 2008 Elsevier Inc. All rights reserved.

1. Introduction

There is a strong need for integrating diverse types of data from different sources to support basic, clinical, translational research for human diseases and this continues to be a significant challenge for the research community. To demonstrate how to address this challenge we use a variety of technologies including ontology, text mining, and visual query to integrate and navigate information in the context of the infectious disease, dengue.

Dengue, one of the most prevalent viral diseases of humans, continues to infect millions of people in the tropical and subtropical regions annually. Dengue virus (DENV) belongs to genus Flavivirus of the family Flaviviridae [1–3] and is transmitted to humans by mosquitoes Aedes aegypti and Aedes albopictus, also known as the Asian tiger mosquito [4,5]. DENV is categorized into four antigenically distinct serotypes [6]: DENV-1, DENV-2, DENV-3, and DENV-4. Currently, all serotypes are co-circulating globally affecting more than 100 countries where over 2 billion people are estimated at risk of infection [7]. Infection by dengue virus can lead to a spectrum of illnesses: dengue fever (DF), dengue hemorrhagic fever (DHF), and dengue shock syndrome (DSS). The most common and least severe form of dengue is DF. Systemic and ocular manifestations associated with DF are fever, nausea, abdominal and joint pain, headaches, double vision, foreign body sensation, and loss of vision. The more severe forms of dengue, DHF and DSS, are rare complications that can be fatal. While all four viral serotypes can cause the disease, infection with one serotype only provides life long immunity to that serotype and does not protect from infection by other serotypes [8]. The first known epidemic of DHF was reported in Manila, Philippines, in 1953–1954. Epidemic activity was re-established in the Pacific Islands and Americas in 1970s, and has since intensified [9].

Finding a vaccine or therapeutic drug for dengue has posed a significant challenge for the research community. While therapeutic-based approaches have been proposed as an alternative to vaccines [10] there are currently no approved vaccines or therapeutic drugs. Elimination of the mosquito vectors is currently the only measure to prevent and control transmission of dengue [11].

There has been a consistent increase in the amount of dengue-related information being generated by research consortia e.g. the Paediatric Dengue Vaccine Initiative (DPVI), and in corporate sector research e.g. the Novartis Institute for Tropical Diseases (NITD). In particular, the production and availability of full length genome sequences of clinical DENV isolates prompted NITD and others to develop the DengueInfo database (www.dengueinfo.org) [12]. While it serves as an important starting point, the database falls short of being an exhaustive repository of such information since it is limited to full genomes and the manual curation required to incorpo-
rate sub-genomic sequences and annotations into the database has not materialized. However, since the establishment of DengueInfo greater efforts to sequence viral genomes have emerged, notably at the Microbial Sequencing Centre of the National Institute of Allergy and Infectious Diseases (NIAID) [13] which is involved in generating sequence data for over 3000 dengue genomes. This initiative has internally standardized clinical data annotated to each genome across all partnering institutions and has standardized its methods of sequence data generation and gene annotation. Curation is provided by the NIAID's Viral Bioinformatic Resource Center and the data is made available through the Broad Institute's dengue database [14].

Other types of dengue-related information are also being generated as a result of several comprehensive clinical studies [15–17] and the types of data being collected (e.g. mouse model survival times or patient responses to vaccine formulations) are diverse, albeit from proprietary sources. Technology integration initiatives are also crucial in supporting clinical and epidemiological studies, for example, in the characterization of dengue transmission, vaccine safety research, and in vaccine trials [18]. These efforts benefit from streamlined information workflows and data accessibility, enhanced data quality and quality control procedures, as well as reduced operational costs.

Despite such initiatives, standardization across public datasets and data types has yet to emerge. The lack of standardization means that finding specific information across research consortia is very difficult. For example, it is currently impossible to reliably find defined groups of sequences such as the NSS protein sequence of all DENV-3 viruses isolated in Thailand that caused DHF.

There are numerous well known hurdles to all forms of data integration and knowledge navigation that are also relevant in the dengue context. These include the data storage mechanisms deployed (e.g. typically only manually accessible through web-based forms facilitating structured query language (SQL) queries against a single or multiple relational databases), the complexity of access protocols for information access and remote transfer [19] and the existence of inconsistent standards for representation and cross mapping of metadata. Critically, the establishment of domain-specific knowledge representations mapped with standardized metadata for semantically related datasets is necessary to facilitate data integration and interoperability initiatives [20]. In some cases, sharing of information between researchers and clinicians is further complicated by confidentiality, ethics and security issues.

Also relevant here is the observation that digital information available as semi-structured (e.g. XML) and unstructured data is estimated to comprise of 53% of all available information in enterprises, and little unstructured (natural language text) or semi-structured data makes its way into data warehouses [21]. This predicament is in part the foundation stone of biomedical text mining initiatives, none of which have focussed on dengue literature [22–24]. Whilst being the most comprehensive information resource for dengue information the corpus of peer reviewed dengue literature is relatively small, PubMed lists approximately 6000 resources for dengue information the corpus of peer reviewed dengue literature [19] and the existence of inconsistent standards for representation and cross mapping of metadata. Critically, the establishment of domain-specific knowledge representations mapped with standardized metadata for semantically related datasets is necessary to facilitate data integration and interoperability initiatives [20]. In some cases, sharing of information between researchers and clinicians is further complicated by confidentiality, ethics and security issues.

Also relevant here is the observation that digital information available as semi-structured (e.g. XML) and unstructured data is estimated to comprise of 53% of all available information in enterprises, and little unstructured (natural language text) or semi-structured data makes its way into data warehouses [21]. This predicament is in part the foundation stone of biomedical text mining initiatives, none of which have focussed on dengue literature [22–24]. Whilst being the most comprehensive information resource for dengue information the corpus of peer reviewed dengue literature is relatively small, PubMed lists approximately 6000 abstracts that contain the word ‘dengue’. To date we know of no initiatives to simplify searching and knowledge discovery of this material by defining restricted vocabularies or by using domain-specific metadata.

In parallel to the research community, which primarily consumes research articles, the disease epidemiology community mainly consumes content available in newsfeeds and blogs for outbreak monitoring, decision support and intervention. These studies require time sensitive information which may also originate from multilingual sources [25]. Both dengue research and epidemiology demand uninhibited access to the un-ordered dengue information distributed across the web and in scientific literature databases.

In this paper, we present a literature driven, ontology-centric navigation infrastructure suitable for deployment in a multiple application domains. Here we specifically illustrate its customization and use in the context of dengue. The components of this infrastructure are a dengue ontology, which is to our knowledge the first formal dengue-specific knowledge representation presented to date, generic content acquisition and ontology instantiation pipelines delivering dengue-related sentences derived by dengue-specific text mining, and a generic visual query tool for reasoning over and interrogating the populated dengue ontology.

2. Methods

In this section, we present a description of the ontology construction, the text mining approaches adopted for instantiating the ontology, and the platform for facilitating visual query formulation.

2.1. Information system overview

The knowledge navigation infrastructure consists of a content acquisition engine that drives the delivery of dengue-specific literature. This engine takes user keywords and retrieves full-text research papers from distributed public repositories (e.g. literature databases or patent databases) and converts them to a custom format ready for text mining. A workflow of natural language processing algorithms customized to recognize dengue-specific nomenclature identifies target concepts or keywords and tags individual sentences according to the terms they contain and instantiates them in the dengue ontology. Sentences are instantiated (as A-boxes), using a custom designed Java program, to the dengue ontology’s literature specification (sentence concept) and relations to instances of document metadata are established to facilitate sentence provenance lookup. Each of the designated target concepts/terms found in sentences are added into the ontology as instances of the respective dengue concepts and object properties between the sentence and the instance of target concept are established. The fully instantiated Web Ontology Language (OWL) ontology is reasoned over using the reasoning engine RACER [26] and its A-box query language nRQL [27]. A custom built visual query interface, (Section 2.3.8), facilitates query navigation over instantiated concept hierarchies, object properties and the visualization of datatype properties in the ontology.

2.2. Dengue ontology

Our goal was to take advantage of the combination of the OWL [28] framework with expressive description logics (DLs) without losing computational completeness and decidability of reasoning systems. The dengue ontology was therefore implemented in the OWL-DL language. For ontology building we used the commercially available composer, TopBraid http://www.topbraidcomposer.com, as the knowledge representation editor. During the knowledge acquisition phase of ontology development, we consulted dengue database annotations, texts from the scientific literature, entries within distributed biological databases and domain experts. The resulting baseline ontology consists of 42 dengue classes and 44 properties with a depth of four levels. Much of the knowledge is represented in the object properties (relations) between concepts. The primary goal of the ontology was to serve as a knowledge framework for aggregation and indexing of sentence instances according to domain concepts and as a query model for knowledge navigation tasks. Accordingly, we considered multiple end users of the instantiated ontology and reviewed published
material from various sources in the public domain. In an appraisal of the content available on web pages and 550 dengue-related blogs we identified that online content predominantly focuses on prevention and control, symptoms, and treatment. In contrast, we identified that the majority of scientific literature on dengue is focused on analyzing different dengue serotypes for their characteristics (genotype, variations in individual proteins, disease manifestations, etc.). We decided to conceptualize the ontology to accommodate concepts in both of these areas of dengue-related study.

2.2.1. Modeling dengue information

In the municipal domain public health officers often show interest in tracking the origins and movements of the viral serotypes and their various disease manifestations as epidemic outbreaks. Dengue researchers are interested in the virulence and specificity of a serotype to adults and/or children and the resulting types of disease manifestation. Moreover, geographic locations of outbreaks and observable trends such as oscillations in serotype proliferation are also frequently researched themes important to public policy, prevention and control initiatives. To model this domain, we defined critical concepts for which information is also frequently generated in the media and the public domain. We identified various types of information about epidemics, clinical investigations, and patients. Primarily we defined the class EpidemiCInfo which models information summarizing dengue outbreaks. It is associated with multiple properties (has_year_of_epidemic, has_country_information, has_clinical_information, has_fatality_count, has_region_information, and has_virus_description) to represent detailed information on past epidemics. ClinicalInformation represents clinical investigations on dengue infected patients. Review of such information allows standardization of clinical criteria for dengue case classification, hospitalization, clinical management tasks, as well as social policy and vector control initiatives. Clinical information can also hold information identifying clinical dengue markers in a population with a unique history of dengue virus epidemics [29]. PatientInformation is modeled to accommodate sources of information which provide detail on dengue infected patients, including age, race, and symptoms of individuals and can frequently serve to illustrate the possible consequences of secondary complications associated with dengue (e.g. ocular complications after infection). Such studies are often carried out by assessing demographic characteristics, ocular symptoms, laboratory parameters, and the severity of dengue in cohorts of patients [30]. Accordingly, we modeled the concept Symptom and two sub-concepts, OcularSymptom and SystemicSymptom, which are related to DengueDiseaseManifestation by the property has_symptom.

The biogeography of epidemics requires the modeling of GeographicInformation which has a sub-hierarchy consisting of Region and Country. Dengue prevention measures adopted by health centers and institutes around the world are modeled through the concept PreventionAndControlSpecification. Instances of such organizations are modeled through ContactInformation concept. Virus information is modeled under VirusDescription and has three sub-concepts, VirusFamily, VirusGenus, and VirusName. The VirusDescription is related to the concepts EpidemiCInformation, Serotype, and MosquitoVector via property has_virus_description. Characterization of the Genotype of a dengue virus, using phylogeny techniques, can facilitate the analysis of lineage diversification for a dengue serotype over a specific time period. The concepts Serotype and ProteinDomain are modeled to facilitate the exploration of relationships among instances of serotype, genotype or sequence descriptions.

The ontology model is also designed to support the knowledge framework for effective text mining of dengue-related information with a DocumentSpecification, which has the following sub-concepts: BlogItem, DocumentIdentifier, NewsfeedItem, PubMedArticle, and Sentence. The full DocumentSpecification becomes instantiated with information provided by the text mining pipeline and ensures that provenance information about instantiated sentences can be derived. So as to anticipate the deployment of the ontology for integration of external databases the DengueDatabaseSpecification class was introduced. It has the sub-class RecordIdentifier under which GeneBankIdentifier, NCBIIdentifier, and PubMedIdentifier are modeled as sub-concepts. Fig. 1 shows the current dengue conceptualization. Our ongoing work includes the expansion of the ontology with more granular domain-specific terms and merging the ontology with formal upper level bio-ontologies.

Within our system, the dengue conceptualization fulfills multiple roles. It explicitly represents the important concepts, both hierarchically and with defined relationships between the domain concepts. The OWL-DL formalism supports the A-box instantiation for storage of instances and their subsequent query with description logic reasoners. For end user navigation of instances, the ontology serves as a query model with rich domain semantics.

2.3. Text mining and ontology population workflow

In this section, we describe the content acquisition, natural language processing, and ontology instantiation strategies. Primarily, ontology instances are generated from full-texts provided by the document delivery system using the text mining toolkit called BioText Suite, (http://datam.i2r.a-star.edu.sg/~kanagasa/BioText/) which performs text processing tasks such as tokenization, part-of-speech tagging, named entity recognition, grounding, and relation mining. An overview of this workflow is shown in Fig. 2.

2.3.1. Content acquisition

We employ a content acquisition engine that takes user keywords and retrieves full-texts from PubMed, Blogs, and Newsfeeds. The engine makes use of Entrez Programming Utilities (http://www.ncbi.nlm.nih.gov/entrez/query/static/eutils_help.html) to retrieve PubMed entries and a web crawler to retrieve content from blogs and newsfeeds. A screen scraping engine is invoked to extract full-texts (when available) and document metadata from retrieved pages. To retrieve PubMed abstracts and articles, we used the query “dengue (serotype OR epidemic)”. Both newsfeeds and blogs (580) were retrieved by conducting a Google search with keywords ‘dengue’ and ‘epidemic’. A subset of 12 blogs and 14 newsfeeds were used to populate the ontology. Samples of newsfeeds and blogs are shown in Table 1. Collections of retrieved web pages are then converted from their original formats (e.g. html, pdf, to ascii text) and made ready for text mining by a customized format converter.

2.3.2. Named entity recognition

We employed a knowledge-based named entity recognition approach using gazetteer lists and rules, since no dengue-specific training corpus was available for use with existing named entity recognition systems (e.g. [31]). Gazetteer processing of full-text documents matches term lists against the text tokens and tags the terms found. During tagging, the corresponding ontology class for the term is added as an attribute, which is used later to guide instantiation to the correct ontology class during population. Separate lists, compiled from different sources including the Wikipedia, Dengueinfo (http://dengueinfo.org) and WHO dengue information site (www.who.int/topics/dengue), are employed for detecting terms such as serotypes, protein domains, country names, etc.
Fig. 1. Graphical illustration of the dengue ontology conceptualization.

Fig. 2. Illustration of the text processing and ontology population workflow carried out by BioTtext (http://research.i2r.a-star.edu.sg/kanagasa/BioText/).
The beginning and ending of a rule set are denoted by `<ConceptName>` and `</ConceptName>`, respectively. A rule set defined for one concept can also be a subset of a rule set defined for another concept.
query formulation ideology, (iii) provision of a translator to convert a graph drawn on the graph canvas to a well-formed and syntactically expressive DL-query language, in our case nRQL, and (iv) the construction and management of a query syntax formulator that enables users to build queries incrementally, giving instantaneous feedback with every single user action. The syntax formulator facilitates complex queries based on multiple triples found in a graph and their connection is based on whether each domain and range in different predicates has similar properties. The syntax formulator also directs both the translation of graph triples into nRQL query atoms triggered by a series of drag ‘n’ drop and joining actions and the submission of the formulated query to the reasoning engine followed by formatting of the results returned.

The KnowleGator query tool receives OWL-DL ontologies as the input and passes them to the reasoner, after which it enters into a dialogue and issues a series of commands to query elementary features of the ontology for visual representation in the components panel of the tool. The navigator interface, Figs. 4 and 5, consists of three main panels, a components panel, the editor panel and the output panel. The components panel renders the ontology as a tree structure showing concepts, roles, and instances. Each concept is pre-queried to retrieve the number of instances it represents and details of object properties are determined. Furthermore, the components panel allows drag and drop functionality for query formulation. The editor panel is structured as a tabbed pane providing rapid switching between groups of functionalities. The ‘Ask a Question’ tab contains the query canvas where questions can be formulated by dragging and dropping an element from the tree structure in the component panel. Each dropped item is associated with an automatically formulated nRQL query. Dragging a single concept invokes the retrieval of all the individuals of a particular concept. Likewise, dragging a named relation (object property) queries for the instances specified in the domain and range of the object property. In the query canvas a complex query is built by extending simpler queries through ‘right click’ enabled instantiated-object property lookup. A separate tab ‘Get the Answer’ shows a query result in tabular form. In the bottom panel the full-text of a sentence is rendered. KnowleGator is a generic tool that can support navigation of any OWL-DL instantiated ontology. It is not restricted to use with the instantiated dengue ontology.

3. Results

3.1. Population performance analysis

We provide a preliminary performance analysis of the text processing and ontology population system in assessing the complete instantiation workflow. This started with a literature search for the query “dengue (serotype OR epidemic)” with our content acquisition engine that retrieved 1405 search results (as of 30 September 2007). Three hundred and twenty-four full-text papers were successfully downloaded. The remaining papers were from journals not subscribed to by our organization, or were abstracts with no downloadable link to the full paper.

Entity recognition, normalization and grounding of the full-text documents occurred at a rate of 3 s/document. After named entity recognition and relation detection, eight documents that matched no instantiation rules were omitted. Ontology instantiation was carried out with the remaining 316 documents. After normalization and grounding, the relation detection identified 13,600 sentences that comprised 387 VirusGenus instances, 1727
DiseaseManifestation instances, 618 MosquitoVector instances, 6543 VirusDescription instances, 178 PhylogenyDescription instances, 3385 Serotype instances, 2122 ProteinDomain instances, 2827 EpidemiologicInformation instances, 2859 GeographicalInformation instances, 766 Symptom instances, and 753 Genotype instances. (It may be noted that a sentence may have two or more instances of different entities.)

3.2. Dengue-related entity recognition

The evaluation of entity recognition shown in Table 3 identified that our text mining achieved near-perfect precision in detecting Serotype and ProteinDomain concepts, which we consider to be due to (i) the named entity recognition relies on well curated dictionaries, and (ii) the ambiguity of these concepts is relatively low in the domain. The high precision for other concepts can also be attributed to (i) above. For the case of the Country concept, our subsequent investigation showed that the precision is affected due to the presence of synonyms for these terms (e.g. America is sometimes referred to as a region). In comparison, the recall is not so impressive since some lexical variants of terms are not recognized, and in some cases because the rule-set failed to recognize the new contexts. In our future work, we plan to make use of advanced entity recognition techniques, e.g. fuzzy term matching and co-reference resolution, and also train our system on larger corpora, to address these issues.

3.3. Occurrence of serotype mentions in the dengue literature

Fig. 6 illustrates the occurrence of serotype mentions in 6000 pubmed abstracts over a 33 year period. Certain trends can be clearly identified, (i) DENV-2 has been consistently receiving more prominent attention than all other serotypes albeit with a strong downtrend in 2002–2005, during which time DENV-1 and 3 had more attention (2004–2005), (ii) mentions of DENV-4 have declined in recent years. As the numbers of mentions are estimates of the true counts, we conducted a statistical analysis to evaluate the reliability of these estimates, and determined the 95% confidence interval for the margin of error in the data points is at most 8 ± 2%. The details of the analysis are described in Appendix A.

3.4. Visual query examples

In this section, we illustrate visual queries to the populated dengue ontology, built using KnowleGator. These queries navigate the knowledge found in these 27 Pubmed documents and 20 web page/blogs using explicitly defined concepts and relations in the ontological query model.

After investigating a number of community-based web sites on dengue, we observed that there is often a keen interest in finding the symptoms associated with different disease manifestations of dengue. As such, we formulated a simple visual query, representing a possible query from the general public, Fig. 4, which specifically illustrates a search for symptoms corresponding to dengue infections from individual serotypes [30]. The concepts VirusDescription, Symptom, DiseaseManifestation, and Sentence are specified. The output identified symptoms for each of the disease manifestations (dengue shock syndrome, dengue hemorrhagic fever and dengue fever) of the dengue virus and the corresponding sentences from which the information was derived, one of which is displayed from the datatype property in the bottom window.

As epidemiological data is often studied to monitor the evolution of a specific viral strain [33] we illustrate a possible query
from a researcher working on dengue-related outbreaks searching for epidemics and the associated serotype, country or the year of epidemic. Fig. 5 illustrates the visual query formulated to find evidence of a dengue epidemic in 1953, specifically sentences describing such an epidemic, the source documents, and the countries in which it occurred are sought. The visual query comprises of the following linked concepts: EpidemicInformation, Country, Sentence, and DocumentIdentifier each with an unspecified variable. An instance of Year, 1953 is selected to restrict the query. The output panel lists four columns of instances bound to each of the concept variables. The content of one sentence stored in an OWL datatype property is shown in the bottom panel and details the DHF epidemic in the Philippines in 1953.

The dengue ontology and the prototype of the visual query platform can be accessed at http://datam1.i2r.a-star.edu.sg/ESTUser-war/userLogin.html.

Fig. 5. Query formulation and search for sentences describing dengue epidemics in 1953.

Table 3
Evaluation of dengue-related entity recognition

<table>
<thead>
<tr>
<th>Concept</th>
<th>No. of mentions in the corpus</th>
<th>Precision</th>
<th>Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>Serotype</td>
<td>114</td>
<td>0.98</td>
<td>0.89</td>
</tr>
<tr>
<td>ProteinDomain</td>
<td>82</td>
<td>1.00</td>
<td>0.74</td>
</tr>
<tr>
<td>Country</td>
<td>29</td>
<td>0.83</td>
<td>0.69</td>
</tr>
<tr>
<td>Region</td>
<td>41</td>
<td>0.89</td>
<td>0.61</td>
</tr>
<tr>
<td>Epidemic year</td>
<td>41</td>
<td>0.88</td>
<td>0.71</td>
</tr>
<tr>
<td>Micro-averaged results</td>
<td></td>
<td>0.95</td>
<td>0.77</td>
</tr>
</tbody>
</table>

Fig. 6. The frequency of normalized mentions of the dengue serotypes (DENV-1, DENV-2, DENV-3, and DENV-4) in PubMed abstracts over the last 33 years. (See Appendix A for an evaluation of the statistical significance of the margin of error in the data points.)
4. Discussion

Biomedical informatics is concerned with the study and application of computational techniques in biological and clinical research, medicine and healthcare. A diversity of computational technologies now support scientists and clinicians in processing, storage and visualization of data and providing decision support, search and data mining capabilities. In point of fact, these domains have meaningful and precise terminologies and complex relations which are often insufficiently captured within infrastructure and software components to facilitate effective navigation over data and content sources.

This is true also for systems supporting dengue informatics. In short there is an ongoing need for expressive semantics to be embedded within knowledge intensive information systems. In line with this need we have sought to develop a novel search infrastructure combining the benefits of complementary language technologies for semantic navigation of the dengue literature. Our infrastructure provides a novel combination of emerging technologies that have not yet received mainstream adoption. This novel combination is able to deliver scientific content derived by automated text mining in a format indexed and query-able according to expressive semantics of a domain-specific ontology. Typically query composition is a bottleneck within life science information systems. This is because form based interfaces or query languages designed for use on relational databases are frequently the only options. In these cases there is limited cognitive support for domain experts and an advanced level of expertise is required to build expressive syntactic queries relevant to the needs of the end user. In our paradigm users are provided with effective cognitive support through a visual query interface facilitating ‘drag and drop’ of concepts and relations from the menu provided by the domain-specific ontology which serves as a semantically rich query model. Moreover queries made to the ontology can be highly complex spanning many relations between objects, without requiring table joins or connection tables to be established, as would be the case with relational databases. Queries using our paradigm can also be more abstract or more granular, using hierarchical and concept terms in addition to instance level keywords to obtain results. Furthermore since text segments are modeled as instances they can be readily queried from the ontology according to the constraints of a visual query issued by the user. The functional relevance of this combination is that to gain new knowledge a user does not have to browse many documents for relevant information; instead they read a subset of sentences selected on the basis of a precise query they have constructed.

Our approach is applicable beyond the scope of our dengue scenario albeit with the customization of specific system components. The components that require customization are the text mining term lists and ontologies. For some domains reusable vocabularies and ontologies can be found in online metadata repositories e.g. Open Biological Ontologies [34] or BioPortal [35]. These can be used with our generic infrastructure components without major modification, namely the knowledge navigation front end. We have previously deployed our approach in the domain of lipidomics [20] and also anticipate community adoption of this new methodology in a range of different applications domains where integrated access to information available only in diverse, legacy, and unstructured formats such as raw text is required. In particular the approach is able to normalize the synonyms found in documents to a consistent data model removing ambiguity of competing terminologies and simplifying user navigation of the literature. It is also relevant to note that sources of information that can be aggregated from a given domain may differ to a great extent in their level of terminology and the relevant content e.g. blogs are different from scientific papers, but the approach can be legitimately applied in both contexts. In summary this approach serves well as a readily customizable way of aggregating content and is a natural evolution to text mining initiatives which have sought to deliver salient information from raw text yet not addressed the needs of users who wish to rapidly navigate the knowledge therein. In addition, when used within a dynamic workflow our system can effectively help scientists keep up to speed with data and literature proliferation.

Having characterized the scope of our approach we now address the specific advantages and limitations of our approach. Our system is not designed as a query–answer engine, and access to information is dependent on the access rights to the source documents that are indexed. Users seeking comprehensive answers to fundamental questions should first pre-select documents for text mining. Moreover, a time delay between document selection and the availability of the knowledge-base for querying must be accommodated by users. These are pragmatic constraints on the usage of the search tool. In spite of this the dengue literature is relatively small, comprising of only 6000 pubmed abstracts. This amount of material, albeit relatively shallow in depth, is readily available with our system. We have also used our pipeline to process other public resources (e.g. full-text open access journals and online content from blogs and newsfeeds). Our pipeline is capable of re-publishing such dengue-related material in a semantically tagged format. In all cases the tagged sentences are accompanied by document provenance so that users can ascertain the sources of information they are interacting with. Researchers can then select the information they deem sufficiently authoritative for their needs.

A central feature of an effective search tool is that it must be simple to learn and easy to use. Query generation and ease of user interaction are key steps in this regard. Our approach provides precise access to pre-indexed fragments of information using a semantically rich query model. Users accustomed to browsing for information found using keyword searches are now at liberty to compose complex queries before browsing subsets of information. This compels the user to construct precise queries with the drag and drop functionality. For this we acknowledge some users experience a learning curve, but do return on subsequent occasions to query the well indexed content of the knowledgebase. Further advantages of the instantiated ontology include the ability to conduct further data mining and trend analysis on defined subsets of instances extracted from dengue literature (Fig. 6).

5. Conclusion

In this paper, we have addressed the need, and presented a solution, for data integration in support of basic, clinical, and translational research into an infectious human disease. We have presented the construction and interrogation of instantiated domain ontology capable of offering unified query access to multiple sources of unstructured information. We have illustrated the system’s ability to aggregate distributed content and semantically index it according to a domain-specific query model. Coupled with an A-box visual query/reasoning tool, the platform facilitates easy access to information that may otherwise only be available in legacy or non-digital data formats. Using this query paradigm we have illustrated that complex queries can be easily constructed using the semantics of the ontology. In the scenario we address, namely dengue, we illustrate legitimate queries that a researcher or health practitioner might choose to ask but not be able to construct using contemporary search engines, form-based online databases or directly from relational databases where multiple tables have to be joined. The domain-specific query model of the ontology and the specified relations between concepts provide an effective query
platform for constructing complex queries. Concept and axiom-based navigation of the indexed sentences generated by text mining is an effective query paradigm. The approach to unstructured information integration that we have adopted is broadly applicable beyond the scope of any particular application domain, albeit with the requirement of some customization to the text mining pipeline and the necessity of a domain-specific ontology. Moreover using this approach it is possible conduct further data analyses on integrated and normalized sets of information.

Acknowledgment

Menaka Rajapakse designed and built the dengue ontology. Rajaraman Kanagasabai engineered the text mining and ontology instantiation pipeline. Wee Tiong Ang designed and built the Knowledge Navigation tool (KnowleGator), Anitha Veeramani implemented the content acquisition engine. Mark Schreiber is a dengue expert in the industrial domain. Christopher Baker was Principal Investigator of this initiative. He conceptualized the information system architecture and coordinated the research and writing of the manuscript.

Appendix A

Before interpreting the graph in Fig. 6, caution must be exercised because the number of occurrences computed is only an estimate of the true count. To judge the reliability of the estimates, we analyzed the margin of error in terms of confidence intervals as follows. We first collected 50 random samples of 25 sentences each from the corpus, using sampling with replacement, and computed the average error of the estimated number of serotype occurrences from the true value in each sample. Performing statistical analysis, we found that the error had a sample mean of $-0.08$ and a standard deviation of 0.01, where the negative sign implies that the actual value is an underestimate of the true value. Thus, assuming normal distribution, we can conclude that the 95% confidence interval for the percentage error will be within 2 standard deviations of the sample mean, or $-0.08 \pm 0.02$. In other words, given a sentence with a serotype mention, our method will miss this sentence 6–10 percent of the time. (Note that this does not contradict the 85% recall in Table 3 because the latter measures the task of identifying all serotype mentions in a sentence from the full-text.) Our main task however is identifying the margin of error in estimating if a PubMed abstract contains a serotype mention. Making use of the above statistics is not straightforward since the sentence distribution in the abstracts is also random. However, as the abstract to be counted will have at least one sentence with a serotype mention, we can provide a worst case confidence interval.


References

[34] http://www.obbfoundry.org/.