Intelligent Information Management, 2016, 8, 66-85
Published Online May 2016 in SciRes. http://www.scirp.org/journal/iim
http://dx.doi.org/10.4236/iim.2016.83006

Predicate Oriented Pattern Analysis for Biomedical Knowledge Discovery

Feichen Shen[1], Hongfang Liu[2], Sunghwan Sohn[2], David W. Larson[3], Yugyung Lee[1]

[1]CSEE Department, University of Missouri, Kansas City, MO, USA
[2]Department of Health Sciences Research, Mayo Clinic College of Medicine, Rochester, MN, USA
[3]Department of Surgery, Mayo Clinic College of Medicine, Rochester, MN, USA
Email: fsm89@mail.umkc.edu

Received 29 February 2016; accepted 28 May 2016; published 31 May 2016

Abstract

In the current biomedical data movement, numerous efforts have been made to convert and normalize a large number of traditional structured and unstructured data (e.g., EHRs, reports) to semi-structured data (e.g., RDF, OWL). With the increasing number of semi-structured data coming into the biomedical community, data integration and knowledge discovery from heterogeneous domains become important research problem. In the application level, detection of related concepts among medical ontologies is an important goal of life science research. It is more crucial to figure out how different concepts are related within a single ontology or across multiple ontologies by analysing predicates in different knowledge bases. However, the world today is one of information explosion, and it is extremely difficult for biomedical researchers to find existing or potential predicates to perform linking among cross domain concepts without any support from schema pattern analysis. Therefore, there is a need for a mechanism to do predicate oriented pattern analysis to partition heterogeneous ontologies into closer small topics and do query generation to discover cross domain knowledge from each topic. In this paper, we present such a model that predicates oriented pattern analysis based on their close relationship and generates a similarity matrix. Based on this similarity matrix, we apply an innovated unsupervised learning algorithm to partition large data sets into smaller and closer topics and generate meaningful queries to fully discover knowledge over a set of interlinked data sources. We have implemented a prototype system named BmQGen and evaluate the proposed model with colorectal surgical cohort from the Mayo Clinic.

Keywords

Biomedical Knowledge Discovery, Pattern Analysis, Predicate, Query Generation

How to cite this paper: Shen, F.C., Liu, H.F., Sohn, S., Larson, D.W. and Lee, Y. (2016) Predicate Oriented Pattern Analysis for Biomedical Knowledge Discovery. *Intelligent Information Management*, **8**, 66-85.
http://dx.doi.org/10.4236/iim.2016.83006

1. Introduction

Researchers and health care practitioners prefer to conduct research in an evidence-based practice by using available research results when making decisions in health care. The main challenge we are facing to support evidence-based research is the big data problem along with large, complex, and dynamic medical data (e.g., clinical research data, EHRs, ontologies). A lot of medical ontologies and tools have been developed for biomedical research and applications. However, these are not sufficient for integrating or mapping unstructured data to structured data that will be significant for evidence-based research. It is mainly due to the lack of the ability to integrate data from such a variety of sources and extract both a cohesive structure and semantics from structured or unstructured data to support evidence-based research.

In order to extract a cohesive structure and semantics, it is essential to know what information exists and what significant relationships are among the related domains (e.g., discover genes responsible for a disease). Semantic Web is able to provide a platform of information exchange for biomedical knowledge. Increasingly, we are also seeing the emergence of biomedical and scientific collaborations. The first notable effort toward connecting scattered medical data is to materialize a data movement by the biomedical community (*i.e.*, Bio2RDF, OBO, LinkedCT) [1]. In addition, the Semantic Web Health Care and Life Sciences Interest Group (HCLSIG) [2] is aimed at utilizing Semantic Web technologies for innovative research and collaboration in the health care and life science domains. In this drive, large amounts of medical data have been specified and shared via machine-readable formats, such as the Resource Description Framework (RDF) and Ontology Web Language (OWL). The ontologies are developed to easily extend the work of others and share across different domains. These Semantic Web technologies make it easier and more practical to integrate, query, and analyze the full scale of relevant biomedical and healthcare data, as well as EHRs for cost effective health care systems [3].

However, to make seamless interoperability and interchange among heterogeneous datasets, there still exist significant difficulties. There are some existing promising semantic approaches for linking different datasets; however, they are computationally expensive and impractical for large scale ontologies since these works may still require human intervention. Furthermore, as the size of data increases drastically, it is difficult to discover information from structured/unstructured data in a single domain or cross domains, especially for those researchers with expertise in a specific domain. Thus, we need to reduce human intervention with the help of process automation in extraction and integration of semantics from structured or unstructured data.

For extraction of a cohesive structure and semantics from structured or unstructured data, identification of meaningful linking, either together within or across a large number of biomedical ontologies, is necessary. vSparQL is introduced to enable application ontologies to be derived from these large, fragmented sources such as the FMA [4]. A series of queries may be generated using large ontologies like the NCI thesaurus by extracting relevant information that is desired for applications [5]. The GLEEN project aims to develop a useful service for simplified, materialized views of complex ontologies [6]. However, these works are limited due to the lack of the comprehensive semantic analysis of large sources and the usage of the knowledge for query processing. We need to connect related information through a reference ontology that becomes a platform to link together multiple ontologies that cover a broad range of related information. Advanced techniques are needed to analyze these larger reference ontologies, rather than simply getting a slice of a reference ontology and applying it for a query process or decision support [5]. There is also some related work on using K-Means and Fuzzy C-Means for clustering microarray data [7] [8], but neither of them are concerned about the semantics of data nor hierarchical clustering.

In this paper, we applied RDF predicate oriented pattern analysis methodology and combined the advantages of Machine Learning with the added rigor of machine-readable semantics in extracting information and generating queries applicable for decision support in clinics [9]-[12]. The approach to converting free text to ontologies is based on text classification methods by identifying related information connected through relationships and classifying them according to their relevance. In addition, queries are evaluated by measuring the content of information, identifying possible extensions or compositions of queries and making a comparison with an existing query benchmark. We have implemented a prototype of the BmQGen system and evaluated the proposed query model based on the predicate oriented clustering with the colorectal surgical cohort from the Mayo Clinic.

2. Methods

The proposed framework BmQGen is based on the following steps: 1) extracting key words from report; 2) converting unstructured (free text) data to semantically structured (RDF/OWL) data; 3) arranging them into groups in a semantically meaningful manner; 4) generating queries for evidence-based practice; 5) providing visualization based query analytics tool.

The main contribution of this paper is knowledge discovery from multiple domains by defining a predicate-oriented model for pattern-based fuzzy clustering, and processing cross-domain queries automatically generated from clustered patterns. **Figure 1** summarizes the proposed system architecture. As seen in this figure, the Free Text Normalizer component first reads unstructured free text documents from doctors' notes and patients' records then uses MedTagger [13] to filter unnecessary terms out and convert free text terms to normalized ones. The Graph Generator component then applies TextRunner [14] on each unstructured normalized document and simplifies the term to generate a RDF triplet. The RDF/OWL data model [15] specifies resources (information on the entities and their relationship in the given document) in the form of triples <subject (S)-predicate (P)-object (O)>, where S denotes the resource, and P denotes aspects of the resource and expresses a relationship between S and O. In the Semantic Cluster component, many such triplets group *n* different ontologies, and these

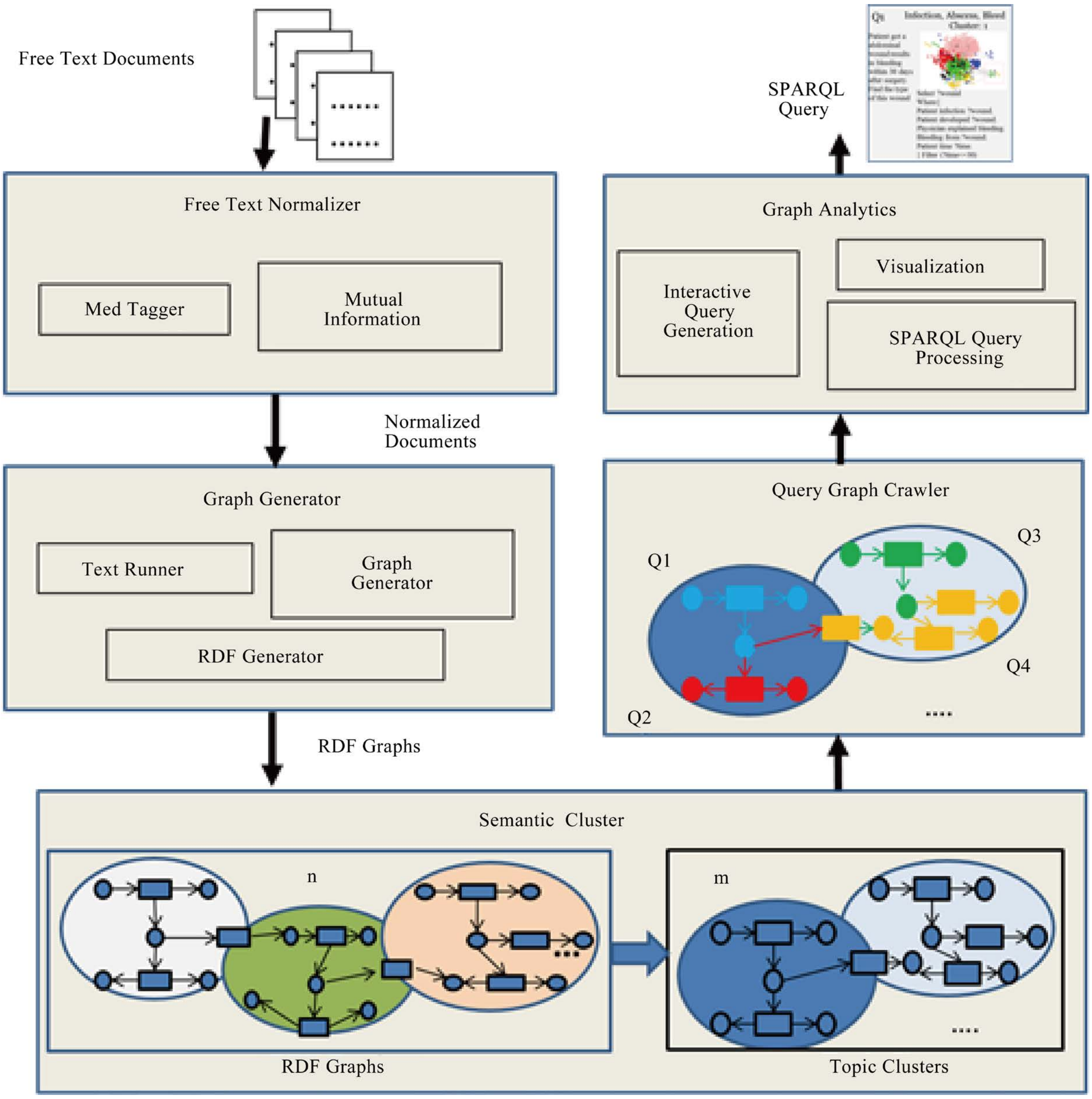


Figure 1. The BmQGen framework.

ontologies can be clustered into *m* semantic groups. The cluster graph is composed of the predicates either from a single domain or different domains. In the Query Graph Crawler component, semantically related queries will be generated for each of the *m* clusters crawling RDF/OWL (predicate) graphs by crawling the cluster graph to generate the query graph with one domain or across domains. The Graph Analytics component allows users to visualize the graphs and generate SPARQL queries through an interactive query interface for selected query graphs.

We first explain how to convert unstructured data to a linked data structure (RDF). We then present a fuzzy clustering algorithm to cluster the RDF/OWL graph. Finally, we evaluate the proposed model through the queries automatically generated from the clusters. The dataflow of the BmQGen framework is shown in **Figure 2**.

The basic steps of the BmQGen framework are described in the remainder of this section.

Step 1: Feature Selection and Concept Annotation.

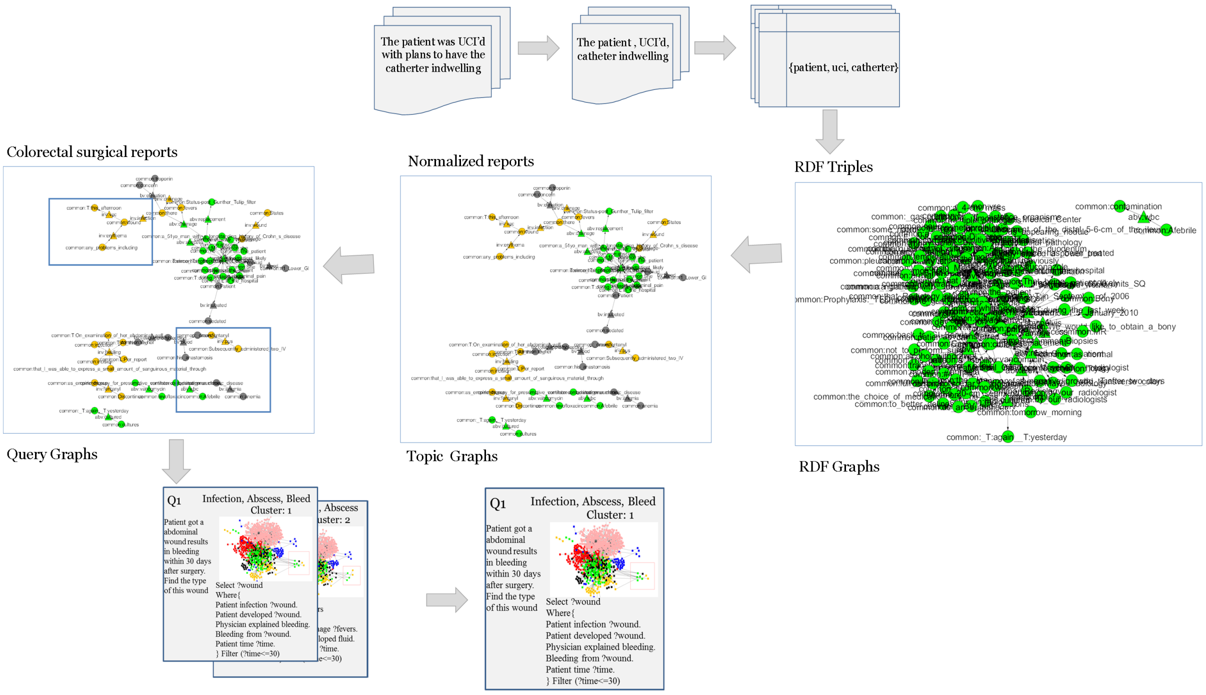
We extracted various free text reports and then performed preprocessing to maintain terms consistently and exclude irrelevant terms, using filtering. Inequality of the likelihood holds between two different values to describe their correlations. To get the inequality of the likelihood between any pairs of words, we then extracted co-occurring terms from free text clinical notes, and calculated the inequality score to cluster them into a different category (domain). In this study, we used MedTagger, which is an open source concept detection and normalization tool through open health natural language processing. Specifically, this tool identifies phrases present in MedLex, a general semantic lexicon created for the clinical domain [16].

The point-wise mutual information was used to assess the inequality of the likelihood for given terms [17].

$$\text{Inequality}(c,o) = \log_2\left(N(c,o) * \log_2 \frac{N(c,o)+0.01}{N(o)} - \log_2 \frac{N(c)}{N}\right.$$

N is the number of observations (e.g., the number of cases for all patients), *N*(*c*) is the number of cases having the concept *c*, *N*(*c*,*o*) is the concept c and the number of cases with a specific complication *o*, and *N*(*o*) is the number of cases with a specific complication *o*.

As the example showed in **Figure 2**, a given input text, "the patient was UCI'd with plans to have the catheter indwelling", MedTagger recognized the uci and catheter concepts and we also found these terms come from ILEUS report. Then we used the point-wise mutual information measurement to calculate any inequality likelihood between these concepts (uci and catheter) and ILEUS. For example, the total number of cases is 1980,



The patient was UCI'd with plans to have the catherter indwelling
The patient , UCI'd, catheter indwelling
{patient, uci, catherter}
RDF Triples
Colorectal surgical reports
Normalized reports
Query Graphs
Topic Graphs
RDF Graphs
Q1
Infection, Abscess, Bleed
Cluster: 1
Cluster: 2

Patient got a abdominal wound results in bleeding within 30 days after surgery. Find the type of this wound

```
Select ?wound
Where{
Patient infection ?wound.
Patient developed ?wound.
Physician explained bleeding.
Bleeding from ?wound.
Patient time ?time.
} Filter (?time<=30)
```

Figure 2. The BmQGen dataflow.

number of ILEUS cases is 400, the number of concept uci among all cases is 600, and the number of concept uci along with ILEUS cases is 500. Based on equation, the inequality between uci and ILEUS is 1.69. We did the same for all other concepts and ranked these concepts by their inequality score and only chose the top 60 of them. Then for any free text which contains top 60 concepts with inequality score, we used OpenIE [18] to get the triple <S-P-O> from the free text annotated with the concepts recognized by MedTagger. For this purpose, we first got a triple {the patient, UCI'd, catherter indwelling}. To make the triple normalized, we looked up the MedTagger concept in the dictionary again and converted the triple to {patient, uci, catherter}. For each complication case, we did the same work above in order to generate six ontologies, respectively.

In **Table 1**, we list some examples about how to map among clinical free text, MedTagger normalized terms and triplets.

Step 2: RDF Graph Construction

First, the top K concepts were selected and each sentence with these concepts in the datasets was annotated with the selected terms. We then extracted the assertions (RDF/OWL triples) from a given free test dataset considering the top K concepts of each domain and generated RDF/OWL triples, respectively. We used OpenIE to extract the triples from the free text. The OpenIE that is based on TextRunner, ReVerb [19] using (PoS) patterns, extracted a significant relation without any relation-specific input. OpenIE used a conditional random field (CRF) classifier to automatically extract triples representing binary relations (Arg1, Relation, Arg2) from sentences. The triples generated from OpenIE were connected to generate RDF graphs.

Step 3: Assertion Clustering and Query Generation

Our assumption for the predicate neighboring patterns (PNP) is that a predicate plays an important role in sharing information and connecting entities among heterogeneous data. For any given domain, the number of unique relations is much less than the number of concepts. Thus, this is another scalable approach for mapping domains than the concept-centric approach. A group of terms (subjects) can be connected to a group of terms (objects) through a single predicate unlike the concept-driven approach. From the unit of <subjects-predicate-objects>, a specific context can be discovered from the associated concepts (subjects, objects). From the neighbors of the predicates, a specific context can be discovered from the association of predicates and their subjects and objects. We can infer/predict missing predicates or missing concepts based on existing contexts. Therefore, we generated a hypothesis that when a graph can be clustered based on PNP patterns, data in the same clustered group have a closer relationship than when in different ones. Predicate neighboring patterns are important to link data together with a variety of domains.

3. Query Generation for Knowledge Discovery

3.1. Predicate Neighboring Patterns

A predicate P is representing a binary relation between two concepts (c1 and c2) in ontology. In RDF/OWL, P is represented as a property to express a kind of relationship (e.g., rdfs:subClassOf) between domain (subject) and range (object). The subject and object can be either from the same ontology or from different ontologies. In our study, relationships are defined by the empirical analysis of ontology data. We are particularly interested in predicates (relationships) that are different from existing approaches like PSPARQL [20] and SPARQLer [21]. Apart from being similar, predicates may share other aspects, e.g., sharing the same subjects or the same objects

Table 1. Mapping among Clinical Free Text, MedTagger terms and RDF triples.

	Clinical Free Text	MedTagger Terms	RDF Triples
1	Patient's abdominal wound was exacerbated by dressing changes	Abdominal wound exacerbated by dressing change	{abdominal_wound, exacerbated_by, dressing_change}
2	Any problems including increased erythema around the wound	Problems including erythema around wound	{problems, erythema, wound}
3	The residual urine levels drop below certain level	Urine drop below level	{urine, drop, below_level}
4	There is substantial further elevation in patient's troponins	Place has further elevation in troponins	{place, elevation, troponins}
5	More hypotension requiring initiation of pressor, to achieve satisfactory blood pressure	Hypotension requiring blood	{hypotension, requiring, blood}

as well as the connectivity between predicates. This forces not only on concepts among graphs but also relationships of the concepts. In this paper, the two types of predicate patterns are defined as follows.

Share Patterns: As **Table 2** shows, this pattern describes the resources sharing relationships between interacting concepts such as shared subjects or shared objects through the given relationship. Assume that two predicates are given as follows: P1 <Si, Oi> and P2 <Sj, Oj> where Si, Sj are a set of subjects and Oi, Oj are a set of objects in given ontologies.

Connection Patterns: According to **Table 3**, the connection pattern is a frequently recurring pattern with predicates observed during query processing as the basis for joining one query pattern to another. This pattern is based mainly on the connectivity of concept(s) through the respective predicates. This type of pattern describes the comprehension of the connectivity relationships between interacting predicates. Assume that two predicates are given as follows: P1 <Si, Oi> and P2 <Sj, Oj> where P1 is directly connected to P2 in the given ontologies Oi, Oj.

We gave the definition of predicate neighboring measurement in our previous paper [9]-[12]. **Figure 3** gives an example to describe how neighboring pattern can be used to measure the closeness among predicates and how similarity matrix can be used to determine the clustering results. **Figure 3(a)** gives a RDF graph with 5 different predicates. In this graph, p1 and p2, p1 and p3, p3 and p4, p4 and p5 are in share patterns. Similarly, p1 and p4, p1 and p5, p2 and p3, p2 and p4, p2 and p5, p3 and p5 are in connection patterns. Based on different patterns each pair of predicates possesses, we calculated their similarity score and build the symmetric similarity matrix as **Figure 3(b)** shows. Then we applied Hierarchical Fuzzy C-Means (HFCM) clustering algorithm on this similarity matrix and built different clusters. As **Figure 3(c)** presents, p1, p2 and p3 are in cluster 1, p3, p4 and p5 are in cluster 2, specifically, p3 is the fuzzy predicate for cluster 1 and cluster 2.

3.2 Hierarchical Fuzzy C-Means Clustering

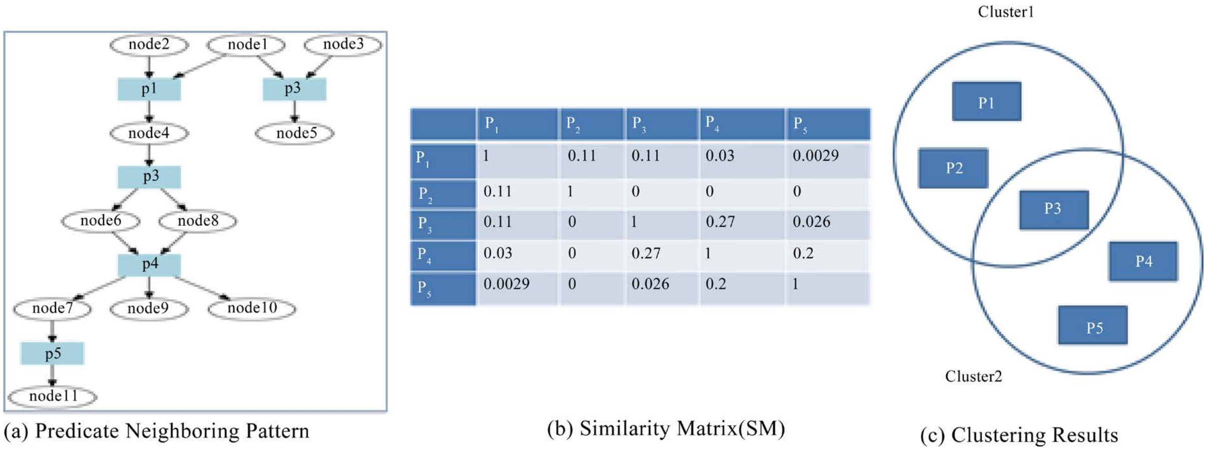
We posited that predicate clustering is a required step for efficient query processing involving the alignment and

Table 2. Predicate sharing patterns.

Patterns	Sharing with S and O through P
Subject-Object Share	Si = = Sj & & Oi = = Oj
Subject Share	Si = = Sj
Object Share	Oi = = Oj

Table 3. Predicate connectivity patterns.

Patterns	Connecting between S and O through P	
	Symbol	Condition
Path Connectivity	Si → P1 → Oi → P2 → Oj	P1 ≠ P2 && Oi = Sj
Cycle Connectivity	Si → P1 → Oi → P2 → Oj	P1 ≠ P2 && Oi = Sj && Si = Oj



	P_1	P_2	P_3	P_4	P_5
P_1	1	0.11	0.11	0.03	0.0029
P_2	0.11	1	0	0	0
P_3	0.11	0	1	0.27	0.026
P_4	0.03	0	0.27	1	0.2
P_5	0.0029	0	0.026	0.2	1

(a) Predicate Neighboring Pattern

(b) Similarity Matrix(SM)

(c) Clustering Results

Figure 3. Predicate neighboring level and weighted similarity.

integration of ontologies. Here we clarify our approach to efficient query processing and query generation within the above theoretical framework. A query processing consists of a collection of several relationships between multiple properties. Given that properties are more closely related to some properties than others, property clustering and partitioning can be utilized for efficient query processing—the task of classifying a collection of properties into clusters. The guiding principle is to minimize inter-cluster similarity and maximize intra-cluster similarity, based on the notion of semantic distance.

To discover the neighboring relation between predicates, we used an innovative Hierarchical Fuzzy C-Means (HFCM) clustering algorithm. We extended a Fuzzy C-Means clustering algorithm [22] with a hierarchical approach applying a heuristic function. In general, we defined a fuzzy hierarchy for clustering setting a machine capacity threshold α to denote a certain number of triplets that each cluster for each level of the hierarchy can hold. The hierarchy can be constructed by applying the Fuzzy C-Means algorithm on each cluster until the number of triplets for each cluster is less than or equal to the threshold α or no further change of numbers of elements for each cluster can be made. To get the optimal number of clusters, we used Silhouette Width to evaluate results and chose the one with the biggest score. The HFCM algorithm is given in **Algorithm 1**.

Algorithm 1. Hierarchical fuzzy C-Means clustering (HFCM).

```
// $P$ is an $n \times n$ predicate similarity matrix, $n$ is the number of predicates in ontologies
// $\delta$ is the threshold of silhouette width $C_{ij}$
Input: $P$, $\delta$
// a hierarchy with a set of clusters $C_{ij}$ the $j^{th}$ cluster at $i^{th}$ level
Output: $C = \{C_{11}, C_{12}, \cdots, C_{ij}\}$
$i$=1
repeat
    // $n$ is the number of input predicates, $m$ is the number of predicates of cluster $j$ at level $i$,
    // optimal $k$ ( $k \le m \le n$ ) from $m$ predicates of the cluster at level $i$ ( $C_i$ ) using $nsw(c_i)$ function
    $sw_1 = nsw(c_i)$ //compute Neighborhood silhouette width
    $k$ = Optimal K( $C_i, sw_1$ )            //fine the optimal $k$ based on $nsw(C_i)$
    Change1 = false
    If (k > 1) then
        Change1 = true
        for $j = 1$ to $k$
            $\mu_{ij} = RM(p_{j1}, p_{j2}, \cdots, p_{jm})$ (// random mean for predicates in $C_{ij}$ (cluster $j$ at level $i$)
        end
        for each $p_{ij} \in P_i$ do
            $\mu_{ij} = Argmin\, w_{ij}^{m}(p_{ij}, \mu_{ij})\ j \in \{1, \cdots, k\}$ //$w_{ij}$ is the degree, m is the fuzzifier
        end
        Change2 = false
        $sw = 0$
        repeat
            for each $\mu_{ij} \in U_i$ do
                Update Cluster( $\mu_{ij}$ )
            end
            for each $p_{ij} \in P_i$ do
                $NCen = Argmin(p_{ij}, \mu_{ij})\ j \in \{1, \cdots, k\}$
                if $NCen \ne \mu_{ij}$ then
                    $\mu_{ij} = NCen$
                    $C_{ij} = C_{ij} \cup p_{ij}$
                    Changed2 = true
                end if
            $sw_2$ = Silhouette Width ( $C_{ij}$ )            //computesil houette width
        while Changed2 == true
while Change1 == true and $sw_2 \ge \delta$
```

3.3. Query Generation

From the HFCM clustering step, predicates are grouped into a number of clusters according to the similarity measurement of the predicate neighboring patterns (PNP). Each cluster will have a graph, called the cluster graph that is composed of the predicates either from a single domain or different domains. The Query Graph Crawler will take over and start crawling the cluster graph to generate the query graph with one domain or across domains. **Figure 4** shows how the Query Graph Crawler generates queries: m clusters are generated after executing the HFCM algorithm and one of them has three predicates, namely uci from the ILEUS domain, transferred from the ABSCESS domain and held from the BLEED domain. As seen in **Figure 4**, we started to generate a query, first visiting the predicate that had the highest in-degree and out-degree and expanded it with its neighbors. In this example, we started with uci and then visited all its neighbors in a descending order of their similarity scores. As the similarity score between uci and held is 0.6, *uci visits held firs*t. And then uci visits transferred with the similarity score 0.5. After all uci's neighbors were visited, we started to select its neighbor that had the highest similarity score. In this example, *held* takes turns to reach its neighbor and find *transferred*. The algorithm runs in an iterative way until there no more neighbors can be visited within the cluster. For example, if a query is generated with a predicate's similarity score t ≥ 0.6, then the generated graph will include only uci and held. Finally, a SPARQL query will be generated by replacing the names of subjects and objects with variables.

4. Results and Discussion

4.1. Specification

The BmQGen prototype system was implemented using Java in an Eclipse Juno Integrated Development Environment [23]. Apache Jena API [24] was used to parse OWL/RDF datasets and retrieve triple information. We used the R computing environment [25] for our experimental validation. We implemented a software plugin for query and schema graph visualization using CytoScape 3.0.2 [26]. In addition, we set the machine capacity threshold α as 1/3 of the total size of the triples.

4.2. Case Study

As a case study, the Mayo Clinic's colorectal surgical reports were used to generate queries by categorizing re-

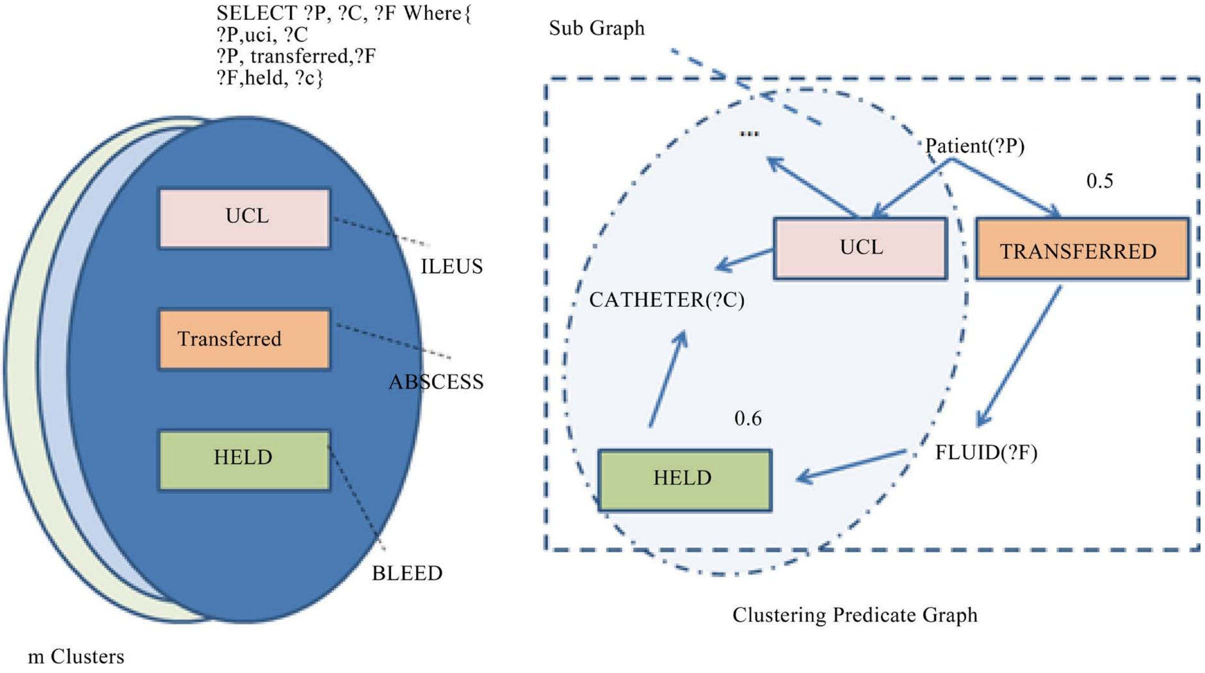


Figure 4. Clustering predicate graph for query generation.

lationships among six colorectal postsurgical complications (deep vein thrombosis/pulmonary embolism, bleeding, wound infections, myocardial infraction, ileus and abscess/leak). Postsurgical complications are related to general or certain type of surgeries. Clinical data of six complications after colorectal surgery were attempted, analyzed to find interesting patterns/associations in a single or multiple complications, and generated comprehensive cross-domain queries that might be useful in conducting evidence-based practice by using available research results. We assume six postsurgical complications represent six domains. Predicate profiles and association patterns are important to link data together with a variety of domains.

4.3. Convert Colorectal Surgical Cohort to Ontology

Our case study has 1980 colorectal surgical cases for 1416 patients between 2005 and 2013 enrolled at the Mayo Clinic in Rochester, MN. We used our previous work, MedTagger to extract concepts from clinical notes written about any complications within 30 days after surgery in cohort. The top 60 concepts (by their inequality scores) were used for information extraction. The definition of the 6 complications is shown in **Table 4**.

We built six ontologies based on the top 60 terms ranked by their inequality scores. **Figure 5** gives visualizetions for each of the six ontologies. We used different colors to indicate different domains, so that ABSCESS is

Table 4. Definition of colorectal postsurgical complication.

Postsurgical Complication	Description
Abscess/Leak (ABSCESS)	An abscess is a painful collection of pus, usually caused by a bacterial infection. Coloanal anastomoses have the highest rates.
Bleeding (BLEED)	Minor and major bleeding is common in anastomotic complications. Epinephrine and saline retention enemas are used to manage serious bleeding. Surgical intervention is necessary if situation is getting worse.
Deep vein thrombosis (DVT)/pulmonary embolism (PE) (DVTPE)	DVT is a condition wherein a blood clot forms in a vein of the deep system. A piece of the clot can break off and travel through the lung, which can cause heart failure, known as PE.
Ileus (ILEUS)	Ileus is defined as bowel obstruction. For small bowel obstruction, 90-100% sensitivity can be achieved by a CT scan of the abdomen and pelvis.
Myocardial infraction (MI)	Myocardial infarction is commonly known as a heart attack. It occurs during surgery or within 30 days after surgery.
Wound infection (INFECTION)	Wound infections commonly present around the fifth postsurgical day and 5-15% of patients have such complication after colorectal surgery.

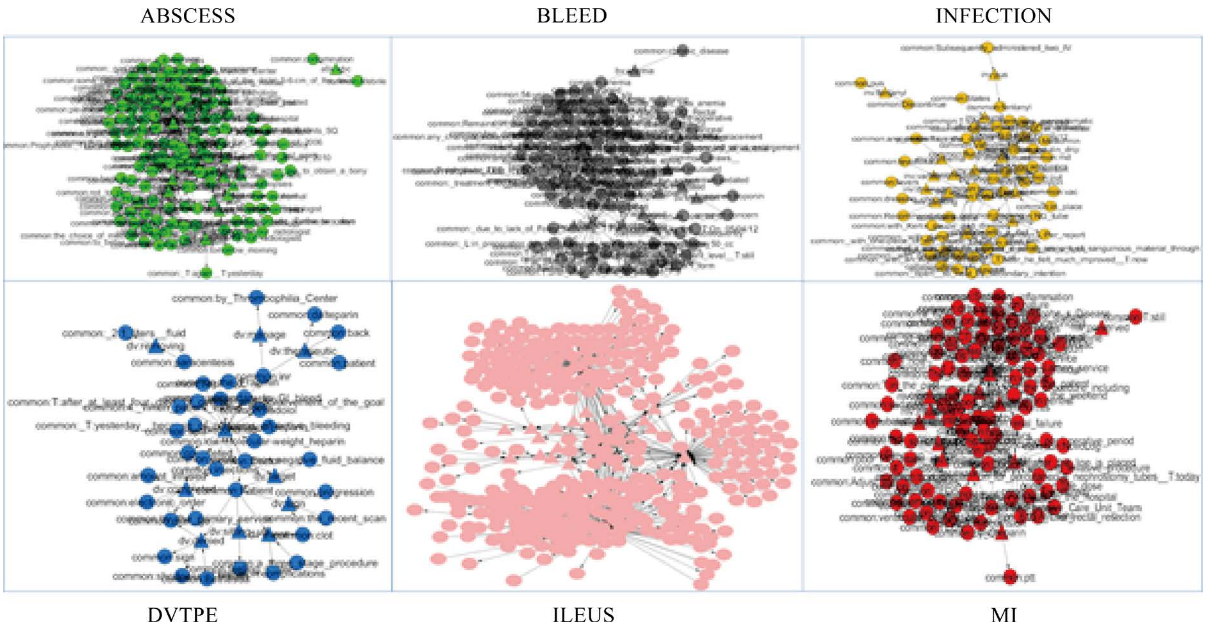


Figure 5. Visualization of 6 complication ontologies.

Table 5. Colorectal surgical cohort.

	# of Subjects	# of Predicates	# of Objects	# of Unique Triples
ABSCESS	63	13	89	220
BLEED	58	13	73	142
DVTPE	19	10	26	32
ILEUS	227	21	204	624
MI	52	12	53	132
INFECTION	26	14	37	60
Total	**445**	**83**	**482**	**1210**

in green, BLEED is in gray, DVTPE is in blue, ILEUS is in pink, MI is in red and INFECTION is in orange. **Table 5** shows the statistics of each ontology. Among the six ontologies, there are a total of 445 subjects, 83 predicates, 482 objects and 1210 triples involved. We then integrated six ontologies together to make them interlinked and prepared to apply a Hierarchical Fuzzy C-Means clustering algorithm on it.

4.4 Hierarchical Fuzzy C-Means (HFCM) Clustering Approach

We applied Hierarchical Fuzzy C-Means (HFCM) clustering for integrated ontology on an input predicate similarity matrix with size 83 × 83. As a result, we got eight different topics. The hierarchical clustering graph is showed in **Figure 6**. The original integrated ontology was partitioned into three intermediate sub-topics based on the optimal Silhouette Width. In addition, three intermediate sub-topics can be further split into eight smaller topics with the best Silhouette Width. Because eight topics cannot be further clustered, then BmQGen stopped the HFCM algorithm and produced eight topics as the final output. We collected the top five predicates for each topic by their total in-degrees and out-degrees and summarized each topic with those predicates. What is more, out of the top five ranked predicates, we also selected top two unique predicates with the most in-degrees and out-degrees for each topic. Unique predicates indicate those predicates that appear in only one topic. Therefore, some topics have unique predicates but some do not. Based on the top predicates and unique predicates, we generated a signature for each topic to summarize the content of each topic.

Figure 7 shows topics 1-4. Topic 1 includes 3 complications (ABSCESS, BLEED and INFECTION) with 24 predicates in total. The top five predicates for Topic 1 are *abv*: *developed*, *inv*: *healing*, *abv*: *drainage*, *bv*: *anemia* and *bv*: *blood*, and the top two unique predicates for Topic 1 are *abv*: *abscess* and *abv*: *replacement*. By analyzing the signature of Topic 1, we found that Topic 1 describes the close relationship among drainage, anemia, blood and incision healing. Similar to Topic 1, Topic 2 also covers three complications (ABSCESS, BLEED and INFECTION) with 16 predicates. Top 5 predicates for Topic 2 are *abv*: *developed*, *bv*: *held*, *inv*: *discontinued*, *inv*: *packed* and *bv*: *drop*, and there are no unique predicates for Topic 2. From this signature, Topic 2 explains that the drop of some life indicators (e.g., hemoglobin) for a patient may be related to the complication (e.g., abscess) developed by such patient; the patient's wound infection is discontinued for the reason that the infection area is packed with gauze. Topic 3 introduces four complications (ABSCESS, BLEED, DVTPE and MI) with 13 predicates. Top 5 predicates for Topic 3 are *mv*: *held*, *bv*: *held*, *mv*: *signs*, *abv*: *drainage* and *bv*: *blood*, and the unique predicates for Topic 3 are none. This signature illustrates BLEED and MI might hold the same symptoms, which are also related to drainage and blood. Topic 4 mentions three different complications (BLEED, DVTPE and ILEUS) with 30 predicates. The top five predicates for Topic 4 are *bv*: *held*, *bv*: *drop*, *ilv*: *clamp*, *ilv*: *fluid* and *ilv*: *dilated*, and the top two unique predicates for Topic 4 are *dv*: *therapeutic* and *ilv*: *bolus*. From this signature, we conclude that fluid has a potential relationship with the dilated, drop of life indicator and ng tube; therapeutic is associated with bolus.

Figure 8 shows topics 5-8. Topic 5 describes 5 complications (ABSCESS, DVTPE, INFECTION, ILEUS and MI) with 23 predicates. The top 5 predicates for Topic 5 are *abv*: *developed*, *inv*: *healing*, *abv*: *drainage*, *bv*: *anemia* and *bv*: *blood*. The top 2 unique predicates for topic 5 are *mv*: *normalized* and *mv*: *aggressive*. The top 5 predicates for Topic 5 convey the exact same information as Topic 1 does. However, unique predicates from Topic 5 tell us that the patient's pain remained poorly-controlled even with an aggressive multimodal; meanwhile, the patient's hypotension had normalized systolic pressure. Topic 6 indicates 4 complications (ABSCESS, BLEED, DVTPE and ILEUS) with 29 predicates. The top 5 predicates for this topic are *ilv*: *ng*, *ilv*: *remove*, *ilv*: *distension*,

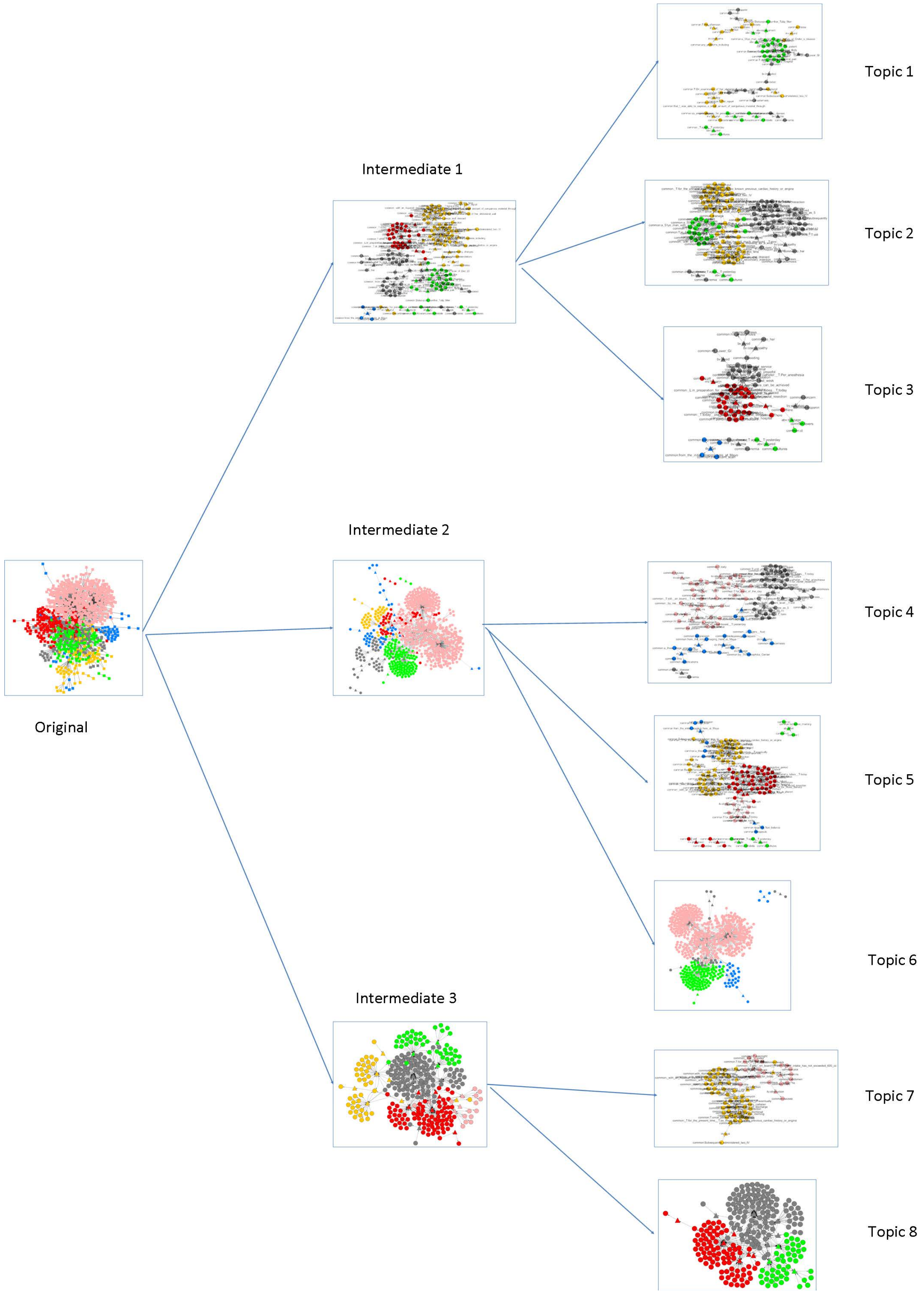
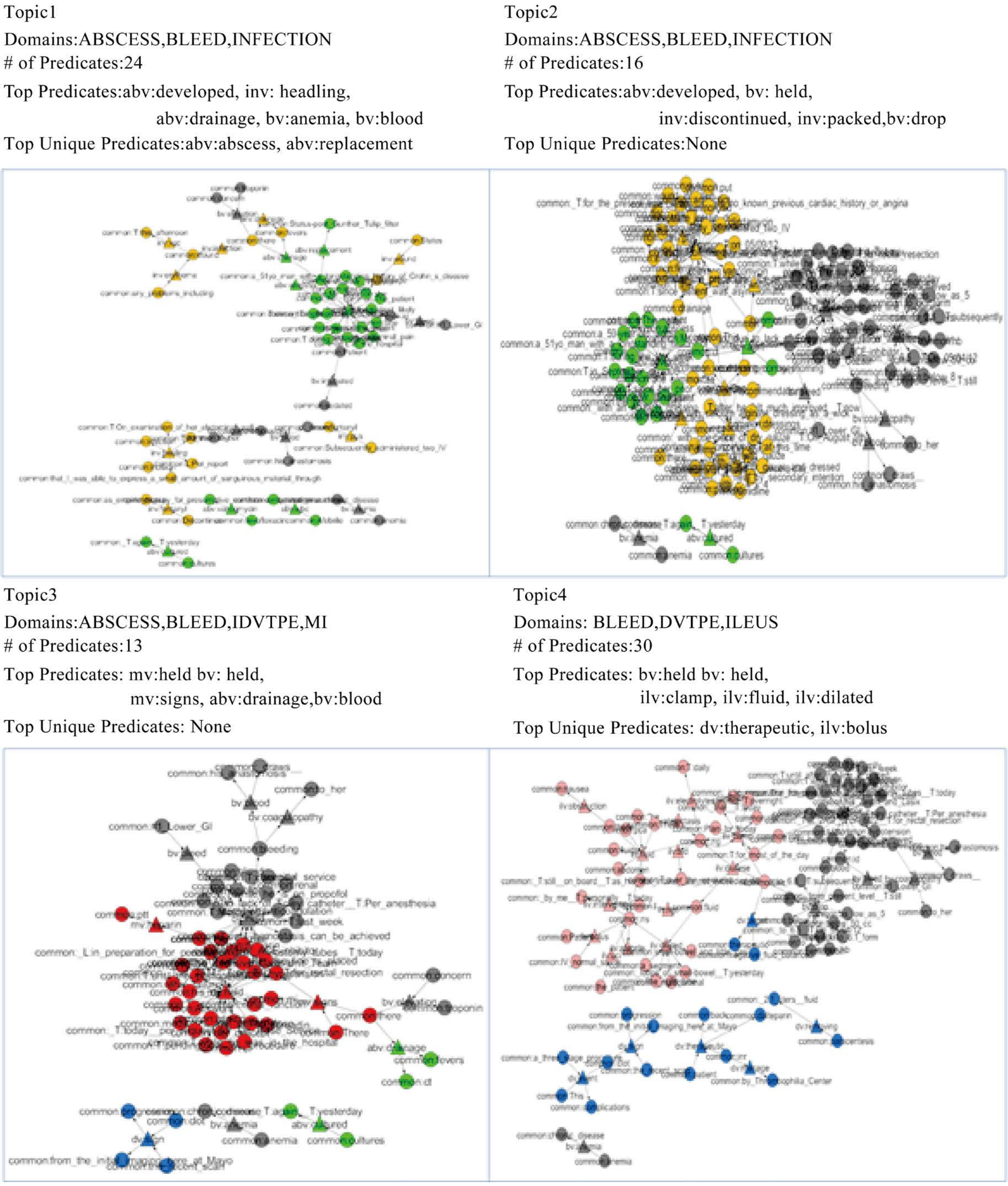


Figure 6. Visualization of hierarchical fuzzy C-Means clustering.



Topic1

Domains:ABSCESS,BLEED,INFECTION
# of Predicates:24

Top Predicates:abv:developed, inv: headling,
abv:drainage, bv:anemia, bv:blood

Top Unique Predicates:abv:abscess, abv:replacement

Topic2

Domains:ABSCESS,BLEED,INFECTION
# of Predicates:16

Top Predicates:abv:developed, bv: held,
inv:discontinued, inv:packed,bv:drop

Top Unique Predicates:None

Topic3

Domains:ABSCESS,BLEED,IDVTPE,MI
# of Predicates:13

Top Predicates: mv:held bv: held,
mv:signs, abv:drainage,bv:blood

Top Unique Predicates: None

Topic4

Domains: BLEED,DVTPE,ILEUS
# of Predicates:30

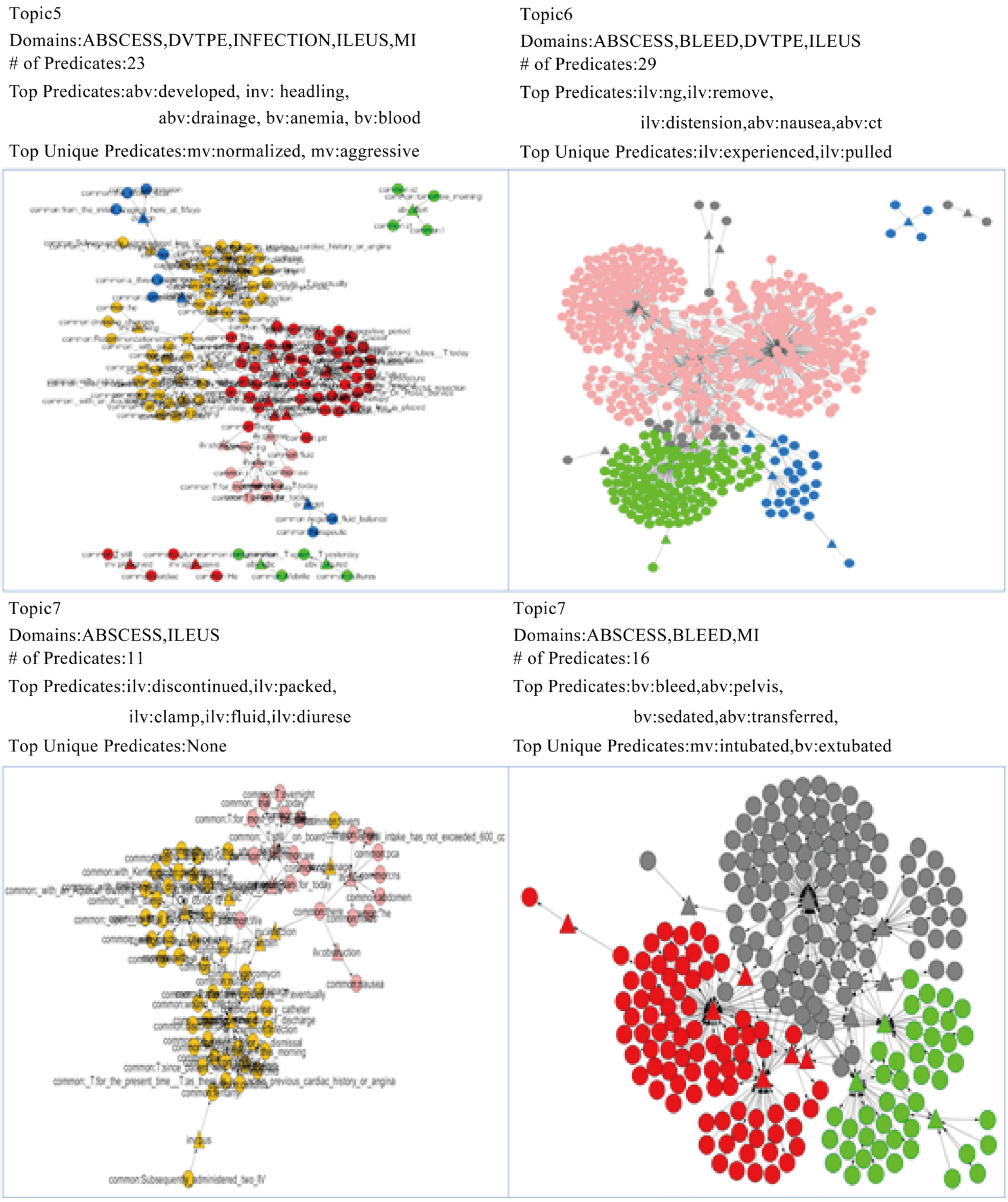
Top Predicates: bv:held bv: held,
ilv:clamp, ilv:fluid, ilv:dilated

Top Unique Predicates: dv:therapeutic, ilv:bolus

Figure 7. Detailed information for Topics 1-4.

abv: *nausea* and *abv*: *ct*. The top 2 unique predicates are *ilv*: *experienced* and *ilv*: *pulled*. This signature summarizes the scenario that a patient's ng tube was pulled out, and this patient also felt nausea and distension. Topic 7 covers 2 complications (ABSCESS and ILEUS) with 11 predicates. The top 5 predicates under Topic 7 are *inv*: *discontinued*, *inv*: *packed*, *ilv*: *clamp*, *ilv*: *fluid* and *ilv*: *diurese*. The top 2 unique predicates are none. This topic is also very similar to Topic 2 and Topic 4 but with more information on diuresis. Topic 8 is related to 3 complications (ABSCESS, BLEED and MI) with 16 predicates. The top 5 predicates involved in this topic are *bv*: *bleed*, *abv*: *pelvis*, *bv*: *sedated*, *abv*: *transferred* and *abv*: *read*. The top 2 predicates are *mv*: *intubated* and *bv*: *extubated*. This topic describes the bleeding situation of the patient's pelvis; such patient was sedated; intubated and extubated operations were also applied on this patient.

We also conducted an experiment among different clustering algorithms to validate that HFCM is the optimal approach to do topic discovery. Silhouette width is a method of validation of consistency within clusters of data. **Figure 9** shows validation for four partitions for each level (one first level and three second levels) with five different clustering algorithms (Hierarchical Fuzzy C-means [22], K-means [27], Clara [28], Pam [29] and Hierarchical Clustering [30]) on the similarity matrix. **Figure 9(a)** shows the splitting from the original ontology to



Topic5
Domains:ABSCESS,DVTPE,INFECTION,ILEUS,MI
# of Predicates:23
Top Predicates:abv:developed, inv: headling, abv:drainage, bv:anemia, bv:blood
Top Unique Predicates:mv:normalized, mv:aggressive

Topic6
Domains:ABSCESS,BLEED,DVTPE,ILEUS
# of Predicates:29
Top Predicates:ilv:ng,ilv:remove, ilv:distension,abv:nausea,abv:ct
Top Unique Predicates:ilv:experienced,ilv:pulled

Topic7
Domains:ABSCESS,ILEUS
# of Predicates:11
Top Predicates:ilv:discontinued,ilv:packed, ilv:clamp,ilv:fluid,ilv:diurese
Top Unique Predicates:None

Topic7
Domains:ABSCESS,BLEED,MI
# of Predicates:16
Top Predicates:bv:bleed,abv:pelvis, bv:sedated,abv:transferred,
Top Unique Predicates:mv:intubated,bv:extubated

Figure 8. Detailed information for Topics 5-8.

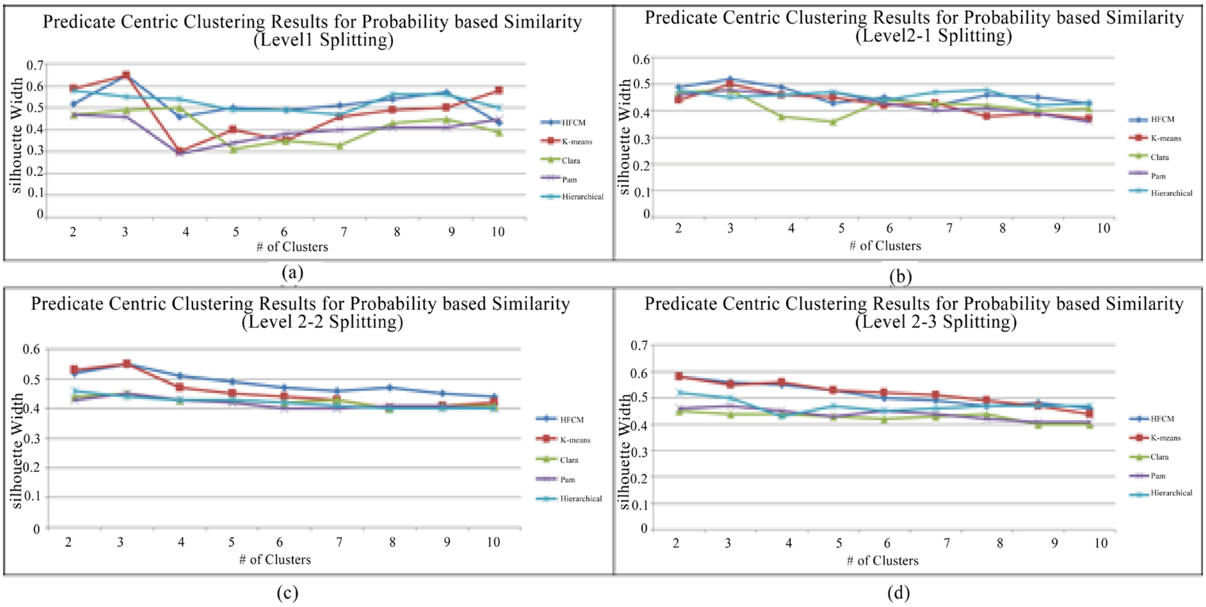


Figure 9. Predicate oriented clustering decision making on different levels.

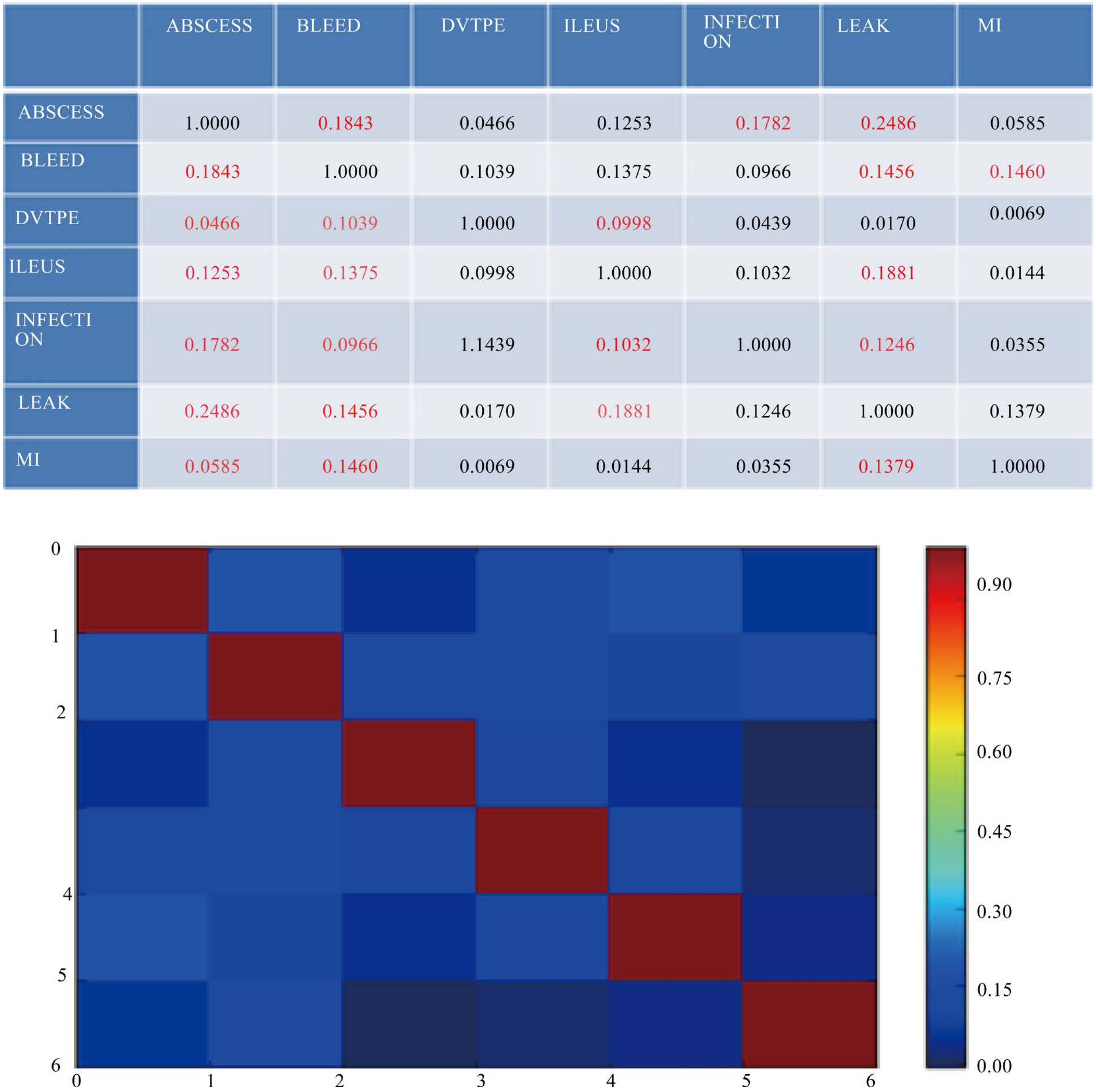
intermediate clusters. Clara, Pam and Hierarchical clustering algorithms showing a relatively stable Silhouette Width for many cases and could not find an optimal cluster number. Both HFCM and K-means give the highest Silhouette Width 0.65 when the cluster number is 3. That shows the reason why the original ontology is split into three intermediate clusters. Similarly, **Figures 9(b)-(d)** present highest Silhouette Width for level 2-1, level 2-2 and level 2-3 splitting, which are 0.52, 0.55 and 0.58 with HFCM, respectively. This explains the reason why the Intermediate 1 Cluster is split into 3 clusters, Intermediate 2 Cluster is split into three clusters and Intermediate 3 Cluster is split into two clusters.

4.5. Validation of Clustering Results with Golden Standard

For those eight generated topics, we used a golden standard file provided by a medical expert, Dr. David W. Larson, in the Colon and Rectal Surgery department at the Mayo Clinics to validate our clustering output. This file lists indications of seven types of complications for 1505 patients after colorectal surgery from 2005 to 2013. In our study, we considered six types of complications by treating ABSCESS and LEAK as the same complication (unlike the golden standard) for the sake of simplicity. A patient may have had no complications or up to seven complications as the golden standard specified. We built correlation metrics based on this benchmark to find out which complications showed a strong positive correlation. **Figure 10** represents the matrices with visualization. The number in red represents the top 3 relative strongest correlations for each complication. It is obvious to see that the complications ABCESS, BLEED and INFECTION have a relative stronger correlation than other complications that verifies that Topic 1 and Topic 2 are valid. ILEUS has a relative stronger relationship with ABSCESS and BLEED, verifying that Topic 6 is valid. LEAK and ILEUS are also strongly associated, verifying that Topic 7 is valid. MI is strongly related to ABSCESS and BLEED, and we can also verify that Topic 8 is valid. DVTPE does not have a very strong relationship with other complications, but this weak correlation with ILEUS, BLEED and ABSCESS is captured by Topics 3, 4, 5 and 6. Therefore, the clusters we generated by the HFCM follow the same correlation provided by the golden standard benchmark.

4.6. Query Generation and Visualization

The SPARQL queries we generated for each cluster are sown in **Figure 11** and **Figure 12**. For the predicate graphs across six domains, we used a rectangle to identify the query boundary out of the whole predicate graph. Queries 1 to 6 are cross domain queries that are automatically generated from each cluster. These queries identify the relationships among different post-surgical complications. For example, INFECTION, ABSCESS and BLEED are closely related to each other through the predicates of *wound*, *bleeding* or *fever*. DVTPE and BLEED are



	ABSCESS	BLEED	DVTPE	ILEUS	INFECTION	LEAK	MI
ABSCESS	1.0000	0.1843	0.0466	0.1253	0.1782	0.2486	0.0585
BLEED	0.1843	1.0000	0.1039	0.1375	0.0966	0.1456	0.1460
DVTPE	0.0466	0.1039	1.0000	0.0998	0.0439	0.0170	0.0069
ILEUS	0.1253	0.1375	0.0998	1.0000	0.1032	0.1881	0.0144
INFECTION	0.1782	0.0966	1.1439	0.1032	1.0000	0.1246	0.0355
LEAK	0.2486	0.1456	0.0170	0.1881	0.1246	1.0000	0.1379
MI	0.0585	0.1460	0.0069	0.0144	0.0355	0.1379	1.0000

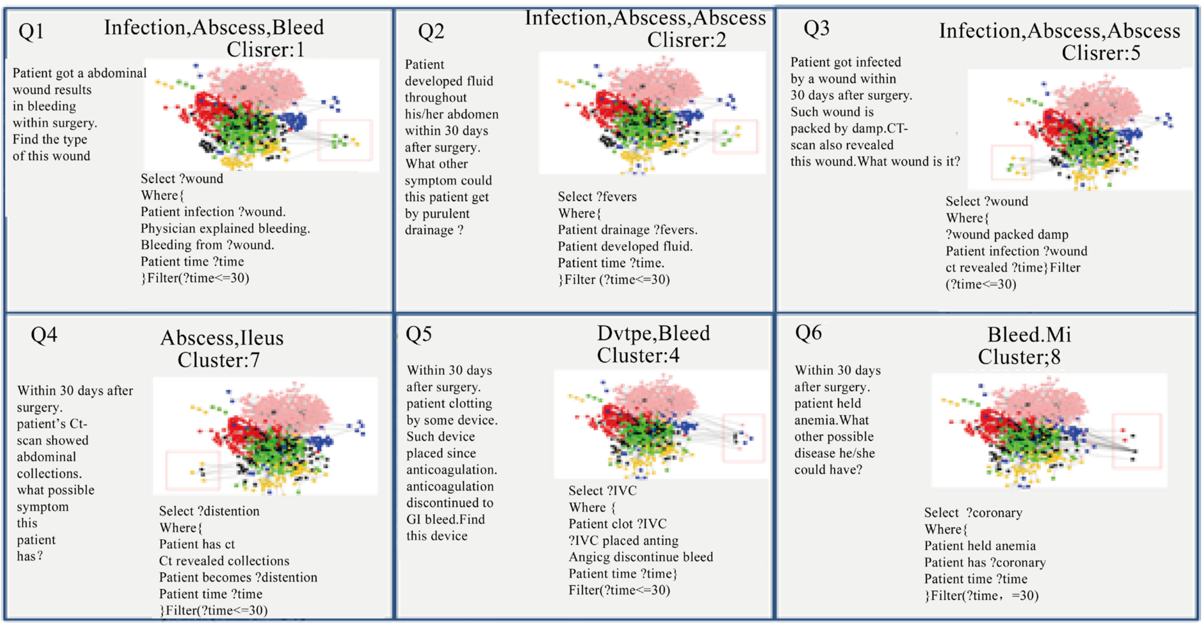
Figure 10. Correlation matrices for golden standard.

usually related through the predicate *clot*. ABSCESS and ILEUS are usually related to each other through *abdominal collections* and *distention*. MI and BLEED are closely related to each other through *anemia* and *coronary*. Queries 7 to 12 are about queries within a single complication. We also find some interesting query patterns for each of the six complications. For instance, in ABSCESS, *sepsis* usually comes with *drainage*. In BLEED, *transfusion* connects to *anemia* and *hemoglobin*. In DVTPE, *coumadin* and *clot* occur together. In ILEUS, *ct scan* and *pelvis* have a close relationship. In INFECTION, *patients* discontinue *wound* after the wound be packed. In MI, pressure and volume overload can be a good treatment for a problem exacerbated by radiation.

4.7. Discussion

Many efforts exist on semantic annotation of data and knowledge discovery on biomedical data [31]-[34]. Especially for [31], the authors annotated a different knowledge domain into their ontology. Moreover, [31] provided a solution to extract existing links among nodes and is capable of predicting potential links or missing links between source and destination. Compared to these works, our research puts more effort on the predicates' relationship detection instead of on the concept nodes.

Furthermore, various data normalization and integration frameworks have been built to complete single domain, as well as cross domain, knowledge discovery. Commonly used medical ontologies are Bio2RDF [35], TMO (Translational Medicine Ontology) [36], Chem2Bio2RDF [37], SIO (Semanticscience Integrated Ontology) [38], ATC (Anatomical Therapeutic Chemical) and DrugBank integration [39], Chem2Bio2OWL [40],



Q1 Infection,Abscess,Bleed
Clisrer:1

Patient got a abdominal wound results in bleeding within surgery. Find the type of this wound

```
Select ?wound
Where{
Patient infection ?wound.
Physician explained bleeding.
Bleeding from ?wound.
Patient time ?time
}Filter(?time<=30)
```

Q2 Infection,Abscess,Abscess
Clisrer:2

Patient developed fluid throughout his/her abdomen within 30 days after surgery. What other symptom could this patient get by purulent drainage ?

```
Select ?fevers
Where{
Patient drainage ?fevers.
Patient developed fluid.
Patient time ?time.
}Filter (?time<=30)
```

Q3 Infection,Abscess,Abscess
Clisrer:5

Patient got infected by a wound within 30 days after surgery. Such wound is packed by damp.CT-scan also revealed this wound.What wound is it?

```
Select ?wound
Where{
?wound packed damp
Patient infection ?wound
ct revealed ?time}Filter
(?time<=30)
```

Q4 Abscess,Ileus
Cluster:7

Within 30 days after surgery. patient's Ct-scan showed abdominal collections. what possible symptom this patient has?

```
Select ?distention
Where{
Patient has ct
Ct revealed collections
Patient becomes ?distention
Patient time ?time
}Filter(?time<=30)
```

Q5 Dvtpe,Bleed
Cluster:4

Within 30 days after surgery. patient clotting by some device. Such device placed since anticoagulation. anticoagulation discontinued to GI bleed.Find this device

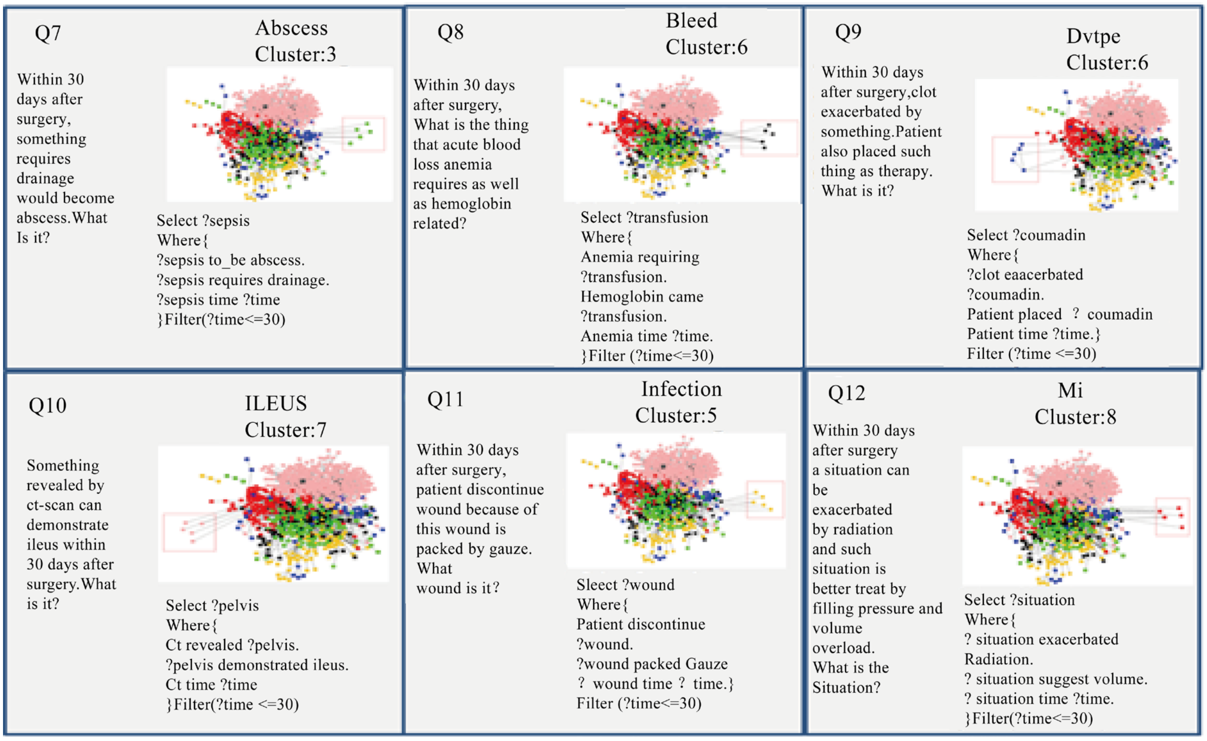
```
Select ?IVC
Where {
Patient clot ?IVC
?IVC placed anting
Angicg discontinue bleed
Patient time ?time}
Filter(?time<=30)
```

Q6 Bleed.Mi
Cluster;8

Within 30 days after surgery. patient held anemia.What other possible disease he/she could have?

```
Select  ?coronary
Where{
Patient held anemia
Patient has ?coronary
Patient time ?time
}Filter(?time，=30)
```

Figure 11. Cross complications queries.



Q7 Abscess
Cluster:3

Within 30 days after surgery, something requires drainage would become abscess.What Is it?

```
Select ?sepsis
Where{
?sepsis to_be abscess.
?sepsis requires drainage.
?sepsis time ?time
}Filter(?time<=30)
```

Q8 Bleed
Cluster:6

Within 30 days after surgery, What is the thing that acute blood loss anemia requires as well as hemoglobin related?

```
Select ?transfusion
Where{
Anemia requiring
?transfusion.
Hemoglobin came
?transfusion.
Anemia time ?time.
}Filter (?time<=30)
```

Q9 Dvtpe
Cluster:6

Within 30 days after surgery,clot exacerbated by something.Patient also placed such thing as therapy. What is it?

```
Select ?coumadin
Where{
?clot eaacerbated
?coumadin.
Patient placed  ?  coumadin
Patient time ?time.}
Filter (?time <=30)
```

Q10 ILEUS
Cluster:7

Something revealed by ct-scan can demonstrate ileus within 30 days after surgery.What is it?

```
Select ?pelvis
Where{
Ct revealed ?pelvis.
?pelvis demonstrated ileus.
Ct time ?time
}Filter(?time <=30)
```

Q11 Infection
Cluster:5

Within 30 days after surgery, patient discontinue wound because of this wound is packed by gauze. What wound is it?

```
Sleect ?wound
Where{
Patient discontinue
?wound.
?wound packed Gauze
?  wound time  ?  time.}
Filter (?time<=30)
```

Q12 Mi
Cluster:8

Within 30 days after surgery a situation can be exacerbated by radiation and such situation is better treat by filling pressure and volume overload. What is the Situation?

```
Select ?situation
Where{
? situation exacerbated
Radiation.
? situation suggest volume.
? situation time ?time.
}Filter(?time<=30)
```

Figure 12. Single complication queries.

Linked Life Data [41], Linked open drug data LODD [42] and LinkedCT: A Linked Data Space for Clinical Trials [43]. These datasets provide a convenient and efficient way for researchers to explore and retrieve valuable information. Similar to these works, we also normalized terms and built our own ontology. However, we used an inequality of likelihood to select top relevant terms and then built ontologies based on these terms.

Meanwhile, a variety of research has been conducted to do systematical and computational knowledge discovery for cross domain knowledge. Specifically, HeteSim [44] and [45] are general systems for relationship extraction and linking detection for heterogeneous datasets. iPHACE [46] investigated a model to acquire knowledge between drug-target interactions. ChemProt [47] provided a database to discover relationships between disease and chemical biology. STITCH 3 [48] retrieved information between chemicals and proteins. In [49], the authors built an integrated platform of drugs, targets and clinical outcomes aimed at supporting drug repurposing. Kinnings *et al*. [50] were able to discover a relationship between drug and disease by deploying chemical systems biology. Campillos *et al*. [51] were able to identify a drug target by using side-effect similarity, and thus found association among the drug, target, and side effect. The Connectivity Map [52] made the use of gene-expression signatures to discover a relationship among small molecules, disease, genes and drugs. In our work, we applied an innovative, unsupervised learning algorithm on a similarity matrix and partitioned the original datasets into several meaningful topics.

Some research focuses on predicate based mining. Shi and Weninger [53] provided a predicate oriented path finding approach to do fact checking in a large knowledge graph. VEPath Cluster [54] proposed a combination of a vertex-centric and edge-centric approach for metapath graph analysis to enhance the clustering quality of cross domain datasets. In addition, pattern analysis was also widely used in data mining and knowledge discovery. In paper [55], the author proposed an interactive way to do data mining by applying pattern mining. In paper [56], the authors used electronic health record data as a use case to introduce an approach to perform data mining and visual analysis on clinical event patterns. WHIDE [57] is a tool for colocation pattern mining in multivariate bio-images. Huang *et al*. [58] accomplished the goal of clinical pathway pattern discovery by using probabilistic topic models. Lasko *et al*. [59] introduced a computational phenotype pattern discovery with unsupervised learning on clinical data. Our research focuses on predicates oriented pattern analysis. In BmQGen, we used our defined predicate neighborhood pattern measurement to quantify the closeness relationship among predicates in RDF.

The limitation in this work is that the datasets we used in this paper were quite noisy and some false positive and true negative cases exist in the mapping between words/phrases and complications. Therefore, the inequality metrics generated by the previous work may not represent the true inequality of the likelihood. What is more, although we built six ontologies, they are mixed with both instances and schemas, so contents within ontologies are not in an abstract level, therefore, clustering results and query outputs will be refined if a more accurately normalized input is given. To provide a better solution, words/phrases identified with a high inequality of likelihood need to be selected to get more normalized data. We also need to try different ontologies and dictionary tools to acquire a better annotated dataset for query generation.

5. Conclusions

In this study, we presented the idea of predicate based pattern analysis, investigated the use of ontology and applied an unsupervised machine learning approach to integrate a heterogeneous unstructured resource with a semi-structured knowledge base. In application level, we achieved specific topic based pattern analysis as well as query generation for cross domain knowledge discovery. A BmQGen framework was proposed to process any RDF/OWL datasets from heterogeneous resources. For the evaluation purpose, we adopted a case study with colorectal postsurgical complications and demonstrated that the BmQGen framework was capable of extracting a cohesive structure and semantics, as well as interesting patterns from structured/unstructured complication datasets. By using the colorectal surgical reports from the Mayo Clinic and golden standard, we successfully validated our clustering results, thereby providing solid evidence for automatic query generation.

In future work, we will improve the accuracy of information extraction by refining the data normalization workflow with different ontologies and extraction/normalization tools (e.g., MetaMap [60]). We will also get more support from domain experts to improve the annotation quality of the datasets. In addition, to improve system performance and scalability, we will utilize a distributed and parallel platform to fulfill efficient clustering and query generation in order to handle big data in biomedical research.

References

[1] Schmachtenberg, M., Bizer, C., Jentzsch, A. and Cyganiak, R. (2014) Linking Open Data Cloud Diagram 2014. http://lod-cloud.net/

[2] Semantic Web Health Care and Life Sciences Interest Group. http://www.w3.org/2001/sw/hcls/

[3] Garde, S., Knaup, P., Hovenga, E.J. and Heard, S. (2007) Towards Semantic Interoperability for Electronic Health Records—Domain Knowledge Governance for Open EHR Archetypes. *Methods of Information in Medicine*, **46**, 332-343.

[4] Shaw, M., *et al*. (2008) Generating Application Ontologies from Reference Ontologies. *AMIA Annual Symposium Proceedings*, **2008**, 672-676.

[5] Nekrutenko, A. and Taylor, J. (2012) Next-Generation Sequencing Data Interpretation: Enhancing Reproducibility and Accessibility. *Nature Reviews Genetics*, **13**, 667-672. http://dx.doi.org/10.1038/nrg3305

[6] Detwiler, L.T., Suciu, D. and Brinkley, J.F. (2008) Regular Paths in Sparql: Querying the Nci Thesaurus. *AMIA Annual Symposium Proceedings*, American Medical Informatics Association.

[7] Sturn, A., Quackenbush, J. and Trajanoski, Z. (2002) Genesis: Cluster Analysis of Microarray Data. *Bioinformatics*, **18**, 207-208. http://dx.doi.org/10.1093/bioinformatics/18.1.207

[8] Dembélé, D. and Kastner, P. (2003) Fuzzy *c*-Means Method for Clustering Microarray Data. *Bioinformatics*, **19**, 973-980. http://dx.doi.org/10.1093/bioinformatics/btg119

[9] MedTQ: Dynamic Topic Discovery and Query Generation for Medical Ontologies. Manuscript Submitted for Publication, 10 January 2016. https://www.dropbox.com/s/9dcpidh5fjdxv1t/MedTQ.pdf?dl=0

[10] Knowledge Discovery from Medical Ontologies in Cross Domains. Manuscript Submitted for Publication, 24 February 2016. https://www.dropbox.com/s/0dzhgqiqneokop0/MedKDD.pdf?dl=0

[11] Shen, F.C. (2015) A Pervasive Framework for Real-Time Activity Patterns of Mobile Users. 2015 *IEEE International Conference on Pervasive Computing and Communication Workshops*, St. Louis, 23-27 March 2015, 248-250.

[12] Shen, F.C., Liu, H.F., Sohn, S.H., Larson, D.W. and Lee, Y. (2015) BmQGen: Biomedical Query Generator for Knowledge Discovery. 2015 *IEEE International Conference on Bioinformatics and Biomedicine* (*BIBM*), Washington DC, 9-12 November 2015, 1092-1097.

[13] MedTagger. http://ohnlp.org/index.php/MedTagger

[14] Yates, A., Cafarella, M., Banko, M., Etzioni, O., Broadhead, M. and Soderland, S. (2007) Textrunner: Open Information Extraction on the Web. *Proceedings of Human Language Technologies*: *The Annual Conference of the North American Chapter of the Association for Computational Linguistics*: *Demonstrations*, Rochester, 22-27 April 2007, 25-26. http://dx.doi.org/10.3115/1614164.1614177

[15] Klyne, G. and Carroll, J.J. (2006) Resource Description Framework (RDF): Concepts and Abstract Syntax.

[16] MedLex. http://www.medlex.com/

[17] Liu, H.F., *et al*. (2014) Facilitating Post-Surgical Complication Detection through Sublanguage Analysis. *Proceedings of the AMIA Joint Summits on Translational Science*, **2014**, 77-82.

[18] Etzioni, O., Fader, A., Christensen, J., Soderland, S. and Mausam (2011) Open Information Extraction: The Second Generation. *Proceedings of the 22nd International Joint Conference on Artificial Intelligence*, **1**, 3-10.

[19] Fader, A., Soderland, S. and Etzioni, O. (2011) Identifying Relations for Open Information Extraction. *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, Edinburgh, 27-29 July 2011, 1535-1545.

[20] Chekol, M.W., Euzenat, J., Genevès, P. and Layaïda, N. (2011) PSPARQL Query Containment. *The* 13*th International Symposium on Database Programming Languages*, 29 August 2011, Seattle, 8 p.

[21] Kochut, K.J. and Janik, M. (2007) SPARQLeR: Extended SPARQL for Semantic Association Discovery. In: Franconi, E., Kifer, M. and May, W., Eds., *The Semantic Web*: *Research and Applications*, Springer, Berlin, 145-159.

[22] Bezdek, J.C., Ehrlich, R. and Full, W. (1984) FCM: The Fuzzy *c*-Means Clustering Algorithm. *Computers & Geosciences*, **10**, 191-203. http://dx.doi.org/10.1016/0098-3004(84)90020-7

[23] Eclipse Juno Integrated Development Environment. https://www.eclipse.org/juno/

[24] McBride, B. (2001) Jena: Implementing the RDF Model and Syntax Specification. *The* 2*nd International Workshop on the Semantic Web*, Hongkong, 1 May 2001.

[25] The R Project for Statistic. http://www.r-project.org/

[26] Shannon, P., *et al*. (2003) Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. *Genome Research*, **13**, 2498-2504. http://dx.doi.org/10.1101/gr.1239303

[27] Hartigan, J.A. and Wong, M.A. (1979) Algorithm AS 136: A K-Means Clustering Algorithm. *Journal of the Royal*

Statistical Society. *Series C* (*Applied Statistics*), **28**, 100-108. http://dx.doi.org/10.2307/2346830

[28] Ng, R.T. and Han, J.W. (2002) CLARANS: A Method for Clustering Objects for Spatial Data Mining. *IEEE Transactions on Knowledge and Data Engineering*, **14**, 1003-1016. http://dx.doi.org/10.1109/TKDE.2002.1033770

[29] Kaufman, L. and Rousseeuw, P.J. (1990) Partitioning around Medoids (Program PAM). In: Kaufman, L. and Rousseeuw, P.J., Eds., *Finding Groups in Data*: *An Introduction to Cluster Analysis*, John Wiley & Sons, Inc., Hoboken, 68-125.

[30] Johnson, S.C. (1967) Hierarchical Clustering Schemes. *Psychometrika*, **32**, 241-254. http://dx.doi.org/10.1007/BF02289588

[31] Chen, B., Ding, Y. and Wild, D.J. (2012) Assessing Drug Target Association Using Semantic Linked Data. *PLoS Computational Biology*, **8**, e1002574.

[32] Doing-Harris, K., Livnat, Y. and Meystre, S. (2015) Automated Concept and Relationship Extraction for the Semi-Automated Ontology Management (SEAM) System. *Journal of Biomedical Semantics*, **6**, 15.

[33] Mate, S., *et al*. (2015) Ontology-Based Data Integration between Clinical and Research Systems. *PloS ONE*, **10**, e0116656.

[34] Willighagen, E.L., *et al*. (2013) The ChEMBL Database as Linked Open Data. *Journal of Cheminformatics*, **5**, 23. http://dx.doi.org/10.1186/1758-2946-5-23

[35] Belleau, F., Nolin, M.-A., Tourigny, N., Rigault, P. and Morissette, J. (2008) Bio2RDF: Towards a Mashup to Build Bioinformatics Knowledge Systems. *Journal of Biomedical Informatics*, **41**, 706-716. http://dx.doi.org/10.1016/j.jbi.2008.03.004

[36] Luciano, J.S., *et al*. (2011) The Translational Medicine Ontology and Knowledge Base: Driving Personalized Medicine by Bridging the Gap between Bench and Bedside. *Journal of Biomedical Semantics*, **2**, S1.

[37] Chen, B., *et al*. (2010) Chem2Bio2RDF: A Semantic Framework for Linking and Data Mining Chemogenomic and Systems Chemical Biology Data. *BMC Bioinformatics*, **11**, 255.

[38] Dumontier, M., *et al*. (2014) The Semanticscience Integrated Ontology (SIO) for Biomedical Research and Knowledge Discovery. *Journal of Biomedical Semantics*, **5**, 14. http://dx.doi.org/10.1186/2041-1480-5-14

[39] Croset, S., Hoehndorf, R. and Rebholz-Schuhmann, D. (2012) Integration of the Anatomical Therapeutic Chemical Classification System and DrugBank Using Owl and Text-Mining. *Proceedings of the* 4*th Workshop of the GI Workgroup* "*Ontologies in Biomedicine and Life Sciences*" (*OBML*), Dresden, 27-28 September 2012.

[40] Chen, B., Ding, Y. and Wild, D.J. (2012) Improving Integrative Searching of Systems Chemical Biology Data Using Semantic Annotation. *Journal of Cheminformatics*, **4**, 6. http://dx.doi.org/10.1186/1758-2946-4-6

[41] Momtchev, V., Peychev, D., Primov, T. and Georgiev, G. (2009) Expanding the Pathway and Interaction Knowledge in Linked Life Data. *Proceedings of the* 8*th International Semantic Web Challenge*, Westfields Conference Center, 25-29 October 2009.

[42] Samwald, M., *et al*. (2011) Linked Open Drug Data for Pharmaceutical Research and Development. *Journal of Cheminformatics*, **3**, 19.

[43] Hassanzadeh, O., Kementsietsidis, A., Lim, L., Miller, R.J. and Wang, M. (2009) LinkedCT: A Linked Data Space for Clinical Trials. arXiv:0908.0567.

[44] Shi, C., Kong, X.N., Huang, Y., Yu, P.S. and Wu, B. (2014) Hetesim: A General Framework for Relevance Measure in Heterogeneous Networks. *IEEE Transactions on Knowledge and Data Engineering*, **26**, 2479-2492. http://dx.doi.org/10.1109/TKDE.2013.2297920

[45] Kong, X.N., Zhang, J.W. and Yu, P.S. (2013) Inferring Anchor Links across Multiple Heterogeneous Social Networks. *Proceedings of the* 22*nd ACM international conference on Conference on Information & Knowledge Management*, Burlingame, 27 October-1 November 2013, 179-188. http://dx.doi.org/10.1145/2505515.2505531

[46] Garcia-Serna, R., Ursu, O., Oprea, T.I. and Mestres, J. (2010) IPHACE: Integrative Navigation in Pharmacological Space. *Bioinformatics*, **26**, 985-986. http://dx.doi.org/10.1093/bioinformatics/btq061

[47] Taboureau, O., *et al*. (2010) ChemProt: A Disease Chemical Biology Database. *Nucleic Acids Research*, 1-6.

[48] Kuhn, M., Szklarczyk, D., Franceschini, A., von Mering, C., Jensen, L.J. and Bork, P. (2012) STITCH 3: Zooming in on Protein-Chemical Interactions. *Nucleic Acids Research*, **40**, D876-D880. http://dx.doi.org/10.1093/nar/gkr1011

[49] Oprea, T.I., *et al*. (2011) Associating Drugs, Targets and Clinical Outcomes into an Integrated Network Affords a New Platform for Computer-Aided Drug Repurposing. *Molecular informatics*, **30**, 100-111.

[50] Kinnings, S.L., Liu, N., Buchmeier, N., Tonge, P.J., Xie, L. and Bourne, P.E. (2009) Drug Discovery Using Chemical Systems Biology: Repositioning the Safe Medicine Comtan to Treat Multi-Drug and Extensively Drug Resistant Tuberculosis. *PLoS Computational Biology*, **5**, e1000423.

[51] Campillos, M., Kuhn, M., Gavin, A.-C., Jensen, L.J. and Bork, P. (2008) Drug Target Identification Using Side-Effect Similarity. *Science*, **321**, 263-266. http://dx.doi.org/10.1126/science.1158140

[52] Lamb, J., *et al*. (2006) The Connectivity Map: Using Gene-Expression Signatures to Connect Small Molecules, Genes, and Disease. *Science*, **313**, 1929-1935.

[53] Shi, B. and Weninger, T. (2015) Fact Checking in Large Knowledge Graphs-A Discriminative Predicate Path Mining Approach. arXiv:1510.05911.

[54] Zhou, Y., Liu, L. and Buttler, D. (2015) Integrating Vertex-Centric Clustering with Edge-Centric Clustering for Meta Path Graph Analysis. *Proceedings of the* 21*th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Hilton, 10-13 August 2015, 1563-1572. http://dx.doi.org/10.1145/2783258.2783328

[55] Van Leeuwen, M. (2014) Interactive Data Exploration Using Pattern Mining. In: Holzinger, A. and Jurisica, I., Eds., *Interactive Knowledge Discovery and Data Mining in Biomedical Informatics*, Springer, Berlin, 169-182. http://dx.doi.org/10.1007/978-3-662-43968-5_9

[56] Gotz, D., Wang, F. and Perer, A. (2014) A Methodology for Interactive Mining and Visual Analysis of Clinical Event Patterns Using Electronic Health Record Data. *Journal of Biomedical Informatics*, **48**, 148-159. http://dx.doi.org/10.1016/j.jbi.2014.01.007

[57] Kölling, J., Langenkämper, D., Abouna, S., Khan, M. and Nattkemper, T.W. (2012) WHIDE—A Web Tool for Visual Data Mining Colocation Patterns in Multivariate Bioimages. *Bioinformatics*, **28**, 1143-1150. http://dx.doi.org/10.1093/bioinformatics/bts104

[58] Huang, Z.X., Dong, W., Ji, L., Gan, C.X., Lu, X.D. and Duan, H.L. (2014) Discovery of Clinical Pathway Patterns from Event Logs Using Probabilistic Topic Models. *Journal of Biomedical Informatics*, **47**, 39-57. http://dx.doi.org/10.1016/j.jbi.2013.09.003

[59] Lasko, T.A., Denny, J.C. and Levy, M.A. (2013) Computational Phenotype Discovery Using Unsupervised Feature Learning over Noisy, Sparse, and Irregular Clinical Data. *PloS ONE*, **8**, e66341.

[60] MetaMap. http://metamap.nlm.nih.gov