iExplore: Interactive Browsing and Exploring Biomedical Knowledge

Vinh Nguyen[1], Olivier Bodenreider[2], Jagannathan Srinivasan[3], Todd Minning[4], Thomas Rindflesch[2], Bastien Rance[2], Ramakanth Kavuluru[5], Hima Yalamanchili[1], Krishnaprasad Thirunarayan[1], Satya Sahoo[6], and Amit Sheth[1]

[1] Kno.e.sis Center, Wright State University, Dayton, Ohio
[2] National Library of Medicine, Bethesda, Maryland
[3] One Oracle Drive, Nashua, New Hampshire
[4] Center for Tropical and Emerging Global Diseases, Univeristy of Georgia, Georgia
[5] Division of Biomedical Informatics, University of Kentucky, Lexington, KY
[6] Case Western Reserve University, Cleveland, Ohio

Abstract. We present iExplore, a Semantic Web based application that helps biomedical researchers study and explore biomedical knowledge interactively. iExplore uses the Biomedical Knowledge Repository (BKR), which integrates knowledge from various sources ranging from information extracted from biomedical literature (from PubMed) to many structured vocabularies in the Unified Medical Language System (UMLS). The current version of BKR provides a unified provenance representation for 12 million semantic predications (triples with a predicate connecting a subject and an object) derived from 87 vocabulary families in the UMLS and 14 million predications extracted from 21 million PubMed abstracts. To engage the domain experts in studying and exploring such a comprehensive knowledge base, we developed the iExplore to: 1) visualize and navigate all the possible semantic predications related to concepts of interest, and 2) search for interesting links between concepts. We also provide an authorization mechanism for SPARQL queries generated by the iExplore to support licensed access to UMLS. We demonstrate the use of iExplore in two scenarios: 1) current research in biomedicine, and 2) re-exploration of two previously known literature-based discoveries. iExplore is available at http://knoesis.wright.edu/iExplore.

1 Data Sources

The Biomedical Knowledge Repository (BKR) was developed at the National Library of Medicine (NLM) as an effort to provide a unified view of biomedical resources from multiple knowledge bases. The core vocabulary schema of BKR is the Unified Medical Language System (UMLS) [2], which has three main components: the Metathesaurus, the Semantic Network, and the Lexicon. Another important component of the BKR, the semantic predications, is extracted from PubMed abstracts. The broad coverage of both UMLS and PubMed has impacted many biomedical applications.

The current UMLS, version 2012AA, integrates and normalizes all the biomedical entities from more than 160 sources (e.g. SNOMEDCT, ICD10, MeSH) into 2 million concepts in the Metathesaurus. Among 160 sources, the BKR represents 1.6 million concepts from 87 sources that do not require their own license. The Semantic Network component also provides a consistent categorization of all concepts into 137 semantic types and 15 semantic groups. Besides the concepts in the vocabulary, 87 sources also contribute 12 million semantic predications involving 650 semantic relations. Furthermore, the BKR also includes 14 million semantic predications extracted from 21 million PubMed abstracts involving 54 semantic relations from the semantic network as predicates. This extraction is performed by the SemRep [7] program and has a precision of 83% and recall of 69%.

Capturing the provenance of semantic predications is important for several reasons. Firstly, a semantic predication may be imprecisely identified by SemRep when extracting from a PubMed abstract. Secondly, UMLS predications may be more reliable than PubMed predications. Thus, using provenance information, one can better assess the reliability of semantic predications or resolve conflicts that may arise. The BKR uses a provenance model [8, 9] to represent the source of semantic predications without using blank nodes.

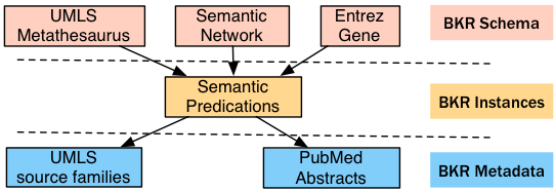

Fig. 1. The Biomedical Knowledge Repository.

2 Design and Implementation

This section discusses the motivation and justification for our design and implementation choices.

2.1 Motivation

Our application was motivated by recent work in the literature-based discovery of new hypotheses. Domain experts manually discovered novel hypotheses such as the link between hypogonadism and diminished sleep quality in aging men [4], or re-composed the Swanson hypotheses [3]. Our main purpose in developing iExplore is to assist the domain experts to explore biomedical knowledge, and generating novel hypotheses semi-automatically [5]. We expect the tool to help domain experts to construct predication subgraphs that may contain interesting candidate links to form new hypotheses. In this paper, we discuss how the domain experts can use their knowledge to guide and drive the knowledge exploration process.

2.2 Design

iExplore is an interactive visualization tool that provides an *intuitive* interface for exploring the BKR. We chose graph based visualization to explore the predications for two reasons. First, graphs are more intuitive for showing directed paths than tabular or other formats. Second, they also naturally capture RDF data as the latter is represented as directed graphs of concepts connected by named relationships.

Predication Subgraphs. We define three basic types of predication subgraphs as building blocks for the visualization. While the subgraph type 1 is useful for finding existing knowledge, the other two subgraph types are useful for finding interesting relationships between concepts. Considering that any two concepts in the BKR may be connected by multiple predications of arbitrary direction, the graphs in Fig. 2 has been simplified for clarity.

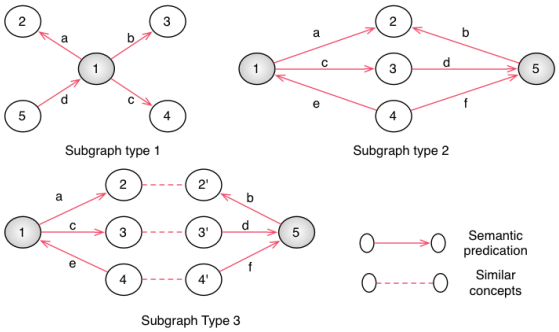

Fig. 2. Three types of predication subgraph.

Subgraph type 1 contains a set of semantic predications related to a central concept. This subgraph allows us to expand a concept to learn the information about it and its neighboring concepts.
Subgraph type 2 contains a set of two directly-connected semantic predications between two given concepts. This subgraph allows us to find the links between two concepts of interest. Here a link is defined as a pair of semantic predications that connect two input concepts through an intermediate concept. Note that a link may not be a path between the two concepts as the direction of the predications to the intermediate concept from the input concepts is arbitrary.
Subgraph type 3 contains a set of two indirectly-connected semantic predications between two given concepts. This subgraph contains two intermediate nodes, each intermediate node is connected to one of the input concepts by a semantic predication. The indirect link (dashed-line) between two intermediate nodes denotes the semantic similarity between them.

Predication subgraph visualization. Given three types of predication subgraphs, the interactive visualization of those subgraphs should meet the following requirements from the BKR.

– Limit the size of the subgraph as large subgraphs cannot fit into a small drawing window. Two approaches can be used to accomplish this. The first approach is "paging", which visualizes large subgraphs, by dividing those

subgraphs into multiple smaller subgraphs and displaying them one by one. The second approach is to find the set of top k links in each subgraph by ranking, as discussed later.
- Expand and collapse subgraphs. On the one hand, expansion should be flexible in selecting concepts to learn or find relationships. On the other hand, collapsing subgraphs keeps the graph compact and readable.
- Filter nodes by predicates or semantic types. As one node may be related to hundreds of other concepts, filtering will help users to quickly navigate to a subset of concepts related to one predicate, or concepts of one specific semantic type.

Ranking. Instead of browsing all the subgraphs to find interesting links, we compute the score of each link and show the most interesting links in each subgraph. The semantic predication links can be ranked using two scores: predication score and semantic similarity score as discussed below.

Predication score. A predication is correct and significant if it has been extracted from a number of PubMed abstracts. Thus counting the number of supporting sources is useful. Furthermore, the less frequent a predication appears, the more interest it holds for biomedical researchers. So we can customize its weight based on its frequency of occurrence in the whole BKR. Thus predication score can be defined analogous to the TF-IDF score in information retrieval where predications are terms and abstracts are documents.

Semantic similarity score. The semantic similarity between two UMLS concepts may be computed based on the transitive closure of UMLS concept hierarchy [1]. However, pre-computation of this score may be computationally expensive because of the large number of UMLS concept pairs (1.6 million by 1.6 million). One way to reduce the computation is to compute similarity between concepts that belong to the same semantic group. This is effective because each concept has a unique semantic group (among the 15 groups) and concepts belonging to different semantic groups are dissimilar.

Features. We provide two main functions to domain expert users: *browsing* existing knowledge and *exploring* hidden relationships.

Browsing. The basic function of iExplore is to help domain experts understand the existing knowledge about concepts by "walking through" all existing predications related to those concepts. Combining interactions such as expanding, collapsing, paging, and filtering of any order within a browsing session would give a good understanding of what information exists for a concept in the BKR.

Exploring. While browsing is useful for a new user to learn about the existing knowledge, exploring may be useful for uncovering implicit links. Our approach for finding and ranking predication subgraphs of type 2 and 3 would help a domain expert see hidden links of their interest, especially, the links containing similar concepts.

Use of semantics. Semantic principles are used extensively in the data sources as well as in the computation performed by iExplore.

Data sources. Biomedical entities sharing the same meaning are gathered into one UMLS concept. Each UMLS concept is categorized by semantic types and

semantic groups. A user may start using iExplore by giving any biomedical term, and the tool will use an UMLS service to find semantically related concepts with their semantic types.

Computation. Semantic types, semantic groups, and UMLS concept hierarchy are very useful. For example, we use 1) semantic types for filtering concepts of interest, and 2) semantic groups for reducing the number of semantic similarity scores to be computed.

2.3 Implementation

This section discusses several practical issues during the implementation of the iExplore. The architecture of iExplore is shown in Fig. 3.

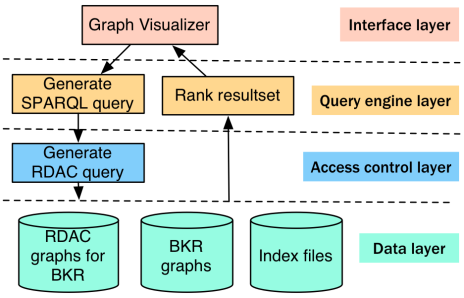

Fig. 3. The iExplore architecture.

Access control. For each interaction in browsing and exploring steps in the interface layer, a set of SPARQL queries are automatically generated. Usually these queries are sent to the triple store containing BKR for execution. However, the UMLS component in the BKR requires an UMLS license to access. Our implementation uses UTS service to authenticate the UMLS licensee, but does not store the credentials. After authentication, we use our approach, called RDAC [6], to validate each SPARQL query. The validation is performed by embedding additional triple patterns for controlling access to each resource returned by the query.

Pre-computation. Given the size of the BKR and the large number of links in the result set, computing the rank on the fly is not practical. Our solution is to pre-compute the TF-IDF scores for each semantic predication and store them in index files. We also pre-compute the semantic similarity scores for each pair of concepts within a semantic group. These scores are queried later for ranking links on the fly.

3 Use case scenarios

We demonstrate the potential of iExplore by providing two scenarios where domain experts can use iExplore in their work.

Searching literature. PubMed has played an important role in biomedical research and it is often consulted while biologists are finding resources from scientific literature. The following sample scenario demonstrates the usage of iExplore in a current research scenario.

Context. Metformin is a widely used drug to treat diabetes but several side effects have been reported in patients taking the drug. One of the most dangerous amongst them is the increase in lactic acid production termed as lactic acidosis. A researcher studying the mechanism of metformin would be interested in exploring the underlying reason behind the increase in lactic acid production with metformin. Assuming that there has not been much research done in this area, one way to approach this is to start reading the related literature in PubMed to build up a hypothesis. Reading all the papers being published on the topic to collect relevant knowledge is time-consuming.

Using iExplore. Here we show how iExplore can be a valuable tool for the domain expert to perform literature-based search and extract all existing relations. For the above example, a quick search on iExplore for the terms metformin and lactic acidosis, resulted in a bunch of relations from literature. Depending on background knowledge of the domain expert and their interest, here are a few picks for them.

Metformin causes **lactic acidosis** which co-exists with **non-insulin dependent diabetes mellitus**.

Metformin causes **autophagy** and **autophagy** affects **lactic acidosis**.

Autophagy augments **ammonia** which produces **lactic acidosis**.

Autophagy affects **hypoxia** which pre-disposes to **lactic acidosis**.

Hypoxia is associated with **angiotensin-II** which produces **ammonia**.

Generate hypothesis. Bridging the gaps between these relations, a simple hypothesis can be built up: "***Metformin*** *causes* ***autophagy*** *which further leads to* ***hypoxia*** *and production of* ***ammonia*** *that pre-disposes to* ***lactic acidosis***". Based on this hypothesis the domain expert can continue the research on how to decrease the pre-disposition to lactic acidosis. For further exploration, a similar search would also provide the relations at molecular level thus saving a lot of time for the domain experts.

Re-exploration scenario. We test our approach by re-exploring the two known discoveries described in [3, 4]. In principle, these hypotheses were generated by investigating the subgraphs and interpreting the links connecting concepts A and C from a set of intermediate nodes B. We had domain experts in our team investigate the two hypotheses and they were able to re-explore the hypotheses easily using iExplore. In each hypothesis, domain experts constructed the subgraphs of their interest after a number of expansion and filtering steps.

4 Challenges and Future Work

Our sample scenario demonstrated the process of how a domain expert can reason across multiple related predications from literature. They use their domain knowledge to quickly gather concepts and relations that suit their reasoning. Such a process is challenging to automatic because incorporating background knowledge into the reasoning process would lead to a combinational explosion in the number of concepts and relations given the size of the BKR. We plan to address these challenging problems in future work.

5 Conclusion

We present our Semantic Web approach to enable engagement of domain experts to study and explore BKR. We described the diversity of data sources that are integrated into the BKR. We explained our design and implementation of iExplore considering the license restriction on the UMLS by NLM. We also provided scenarios to demonstrate how iExplore can be used in biomedical research.

Acknowledgements This research was supported by an appointment to the Research Participation Program at National Library of Medicine, and the NIH R01 Grant number 1R01HL087795-01A1. We thank Amber Mccurdy for help.

References

1. M. Batet, D. Sánchez, and A. Valls. An ontology-based measure to compute semantic similarity in biomedicine. *Journal of Biomedical Informatics*, 44(1):118–125, 2011.
2. O. Bodenreider. The unified medical language system (umls): integrating biomedical terminology. *Nucleic acids research*, 32(suppl 1):D267–D270, 2004.
3. D. Cameron, O. Bodenreider, H. Yalamanchili, T. Danh, S. Vallabhaneni, K. Thirunarayan, A. Sheth, and T. Rindflesch. A graph-based recovery and decomposition of swansons hypothesis using semantic predications. *Journal of Biomedical Informations*, 2012.
4. C. Miller, T. Rindflesch, M. Fiszman, D. Hristovski, D. Shin, G. Rosemblat, H. Zhang, and K. Strohl. A closed literature-based discovery technique finds a mechanistic link between hypogonadism and diminished sleep quality in aging men. *Sleep*, 35(2):279, 2012.
5. V. Nguyen, O. Bodenreider, T. Minning, and A. Sheth. The knowledge-driven exploration of integrated biomedical knowledge sources facilitates the generation of new hypotheses. In *Proceedings of the First International Workshop on Linked Science, ISWC*, 2011.
6. V. Nguyen, R. Kavuluru, O. Bodenreider, and A. Sheth. A resource-based discretionary access control model for linked datasets. *Technical report, Wright State University*, 2011.
7. T. Rindflesch and M. Fiszman. The interaction of domain knowledge and linguistic structure in natural language processing: interpreting hypernymic propositions in biomedical text. *Journal of biomedical informatics*, 36(6):462–477, 2003.
8. S. Sahoo, V. Nguyen, O. Bodenreider, P. Parikh, T. Minning, and A. Sheth. A unified framework for managing provenance information in translational research. *BMC Bioinformatics*, 2011.
9. S. S. Sahoo, O. Bodenreider, P. Hitzler, A. Sheth, and K. Thirunarayan. Provenance context entity (pace): scalable provenance tracking for scientific rdf data. SSDBM'10, pages 461–470, 2010.

Table 1. Appendix: Compliance with SW Challenge Minimal Requirements and Desirable Features

✓ The application has to be an end-user application.
The iExplore has an intuitive graph-based interface for domain experts to interact with at http://knoesis.wright.edu/iExplore/iExplore.html.
✓ The information sources used should be under diverse ownership or control.
The BKR integrates data from multiple sources of different ownerships. The semantic predications extracted from Medline should be publicly available. However, the UMLS is a licensed data source and it requires UTS authentication for each NLM licensee.
✓ The information sources used should be heterogeneous.
The information sources originally are in different file formats (e.g. XML, RDB). The heterogeneity is resolved by converting all data sources into RDF.
✓ The information sources should contain substantial quantities of real world data.
All data in the BKR are from real world data (UMLS and PubMed).
✓ Meaning must be represented using Semantic Web technologies.
The whole BKR is represented in RDF. UMLS class and property hierarchies are presented using RDFS classes and properties.
✓ Data must be processed in interesting ways to derive useful information.
The semantic predications extracted both UMLS and PubMed provide a rich biomedical source for many types of applications as explained in Section 1.
✓ This semantic information processing has to play a central role.
We described the extensive usage of semantics in our data sources, ranking and computation in the Semantics subsection of Section 2.
✓ The application provides an attractive and functional Web interface.
The iExplore has an intuitive graph interface for domain expert users. The end users do not need to know the semantic technologies behind. Click and click!!!
✓ The application should be scalable.
The BKR can be updated on a daily basis from newly published PubMed abstracts. We continue to integrate more data sources from different organisms into the BKR using the approach described in [5].
✓ Functionality is different from or goes beyond pure information retrieval.
The subgraphs queried from iExplore is semantically derived from diverse sources respecting encoded semantics.
✓ The application has clear commercial potential and/or large existing user base.
The broad coverage of UMLS and PubMed in the BKR allows it to be used by all biomedical researchers.
✓ Contextual information is used for ratings or rankings.
Our ranking enhances conventional techniques by using the semantic similarity between concepts and the provenance-based statistics of semantic predications for computing scores as explained in Section 2.
✓ The results should be as accurate as possible (e.g. ranking of results according to context).
By counting the number of PubMed abstracts per predication and capturing the "interesting" predications, we can eliminate erroneous predications in the result set.