Life-iNet: A Structured Network-Based Knowledge Exploration and Analytics System for Life Sciences

Xiang Ren[1], **Jiaming Shen**[1], **Meng Qu**[1], **Xuan Wang**[1], **Zeqiu Wu**[1], **Qi Zhu**[1], **Meng Jiang**[1]
Fangbo Tao[1], **Saurabh Sinha**[1,2], **David Liem**[3], **Peipei Ping**[3], **Richard Weinshilboum**[4], **Jiawei Han**[1]

[1] Department of Computer Science, University of Illinois Urbana-Champaign, IL, USA
[2]Institute of Genomic Biology, University of Illinois at Urbana-Champaign, IL, USA
[3]School of Medicine, University of California, Los Angeles, CA, USA
[4]Department of Pharmacology, Mayo Clinic, MN, USA
[1,2]{xren7, js2, xwang174, mengqu2, zeqiuwu1, qiz3, mjiang89, ftao2, sinhas, hanj}@illinois.edu
[3]{dliem, pping}@mednet.ucla.edu [4]weinshilboum.richard@mayo.edu

Abstract

Search engines running on scientific literature have been widely used by life scientists to find publications related to their research. However, existing search engines in the life-science domain, such as PubMed, have limitations when applied to exploring and analyzing factual knowledge (*e.g.*, disease-gene associations) in massive text corpora. These limitations are mainly due to the problems that factual information exists as an unstructured form in text, and also keyword and MeSH term-based queries cannot effectively imply semantic relations between entities. This demo paper presents the Life-iNet system to address the limitations in existing search engines on facilitating life sciences research. Life-iNet automatically constructs structured networks of factual knowledge from large amounts of background documents, to support efficient exploration of structured factual knowledge in the unstructured literature. It also provides functionalities for finding distinctive entities for given entity types, and generating hypothetical facts to assist literature-based knowledge discovery (*e.g.*, drug target prediction).

1 Introduction

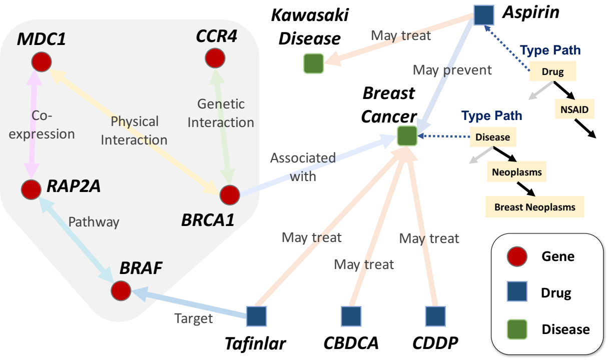


Figure 1: A snapshot of the structured network in Life-iNet.

Scientific literature is an important resource in facilitating life science research, and a primary medium for communicating novel research results. However, even though vast amounts of biomedical textual information are available online (*e.g.*, publications in PubMed, encyclopedic articles in Wikipedia, ontologies on genes, drugs, etc.), there exists only limited support of exploring and analyzing relevant factual knowledge in the massive literature (Tao et al., 2014), or of gaining new insights from the existing factual information (McDonald et al., 2005; Riedel and McCallum, 2011). Users typically search PubMed using keywords and Medical Subject Headings (MeSH) terms, and then rely on Google and external biomedical ontologies for everything else. Such an approach, however, might not work well on capturing different entity relationships (*i.e.*, facts), or identifying publications related to facts of interest.

For example, a biologist who is interested in `cancer` might need to check what specific diseases belong to the category of `breast_neoplasms` (*e.g.*, *breast cancer*) and what genes (*e.g.*, *BRCA1*) and drugs (*e.g.*, *Aspirin*, *Tafinlar*) are related to *breast cancer*, and might need a list of related papers which study and discuss about these disease-gene relations. For cancer experts, they might want to learn about what genes are distinctively associated with `breast_neoplasms` (as compared to other kinds of cancers), whether there exists other genes that are potentially associated with `breast_neoplasms` entities, and whether there exist other drugs that can also treat *breast cancer*.

• **Previous Efforts and Limitations.** In life sciences domain, recent studies (Ernst et al., 2016; Szklarczyk et al., 2014; Thomas et al., 2012;

Kim et al., 2008) rely on biomedical entity information associated with the documents to support entity-centric literature search. Most existing information retrieval systems exploit either the MeSH terms manually annotated for each PubMed article (Kim et al., 2008) or textual mentions of biomedical entities automatically recognized within the documents (Thomas et al., 2012), to capture the entity-document relatedness. Compared with traditional keyword-based systems, current entity-centric retrieval systems can identify and index entity information for documents in a more accurate way (to enable effective literature exploration), but encounter several challenges, as shown below, in supporting exploration and analysis of factual knowledge (*i.e.*, entities and their relationships) in a given corpus.

• **Lack of Factual Structures:** Most existing entity-centric systems compute the document/corpus-level co-occurrence statistics between two biomedical entities to capture the relations between them, but cannot identify the semantic relation types between two entities based on the textual evidence in a specific sentence. For example, in Fig. 1, relations between `gene` entities should be categorized as `CoExpression`, `GeneticInteraction`, `PhysicalInteraction`, `Pathway`, etc. Extracting typed entity relationships from unstructured text corpus enables: (1) structured search over the factual information in the given corpus; (2) fine-grained exploration of the documents at the sentence level; and (3) more accurate identification of entity relationships.

• **Limited Diversity and Coverage:** There exist several biomedical knowledge bases (KBs) (*e.g.*, Gene Ontology, UniProt, STRING (Szklarczyk et al., 2014), Literome (Poon et al., 2014)) that support search and data exploration functionality. However, each of these KBs is highly specialized and covers only a relatively narrow topic within life sciences (Ernst et al., 2016). Also, there is limited inter-linkage between entities in these KBs (*e.g.*, between `drug`, `disease` and `gene` entities). An integrative view on all aspects of life sciences knowledge is still missing. Moreover, many newly emerged entities are not covered in current KBs, as the manual curation process is time-consuming and costly.

• **Restricted Analytic Functionality:** Due to the lack of notion for factual structures, current retrieval and exploration systems have restricted functionality at analyzing entity relationships—they mainly focus on entity-centric literature search (Ernst et al., 2016; Thomas et al., 2012) and exploring entity co-occurrences (Kim et al., 2008). In practice, analytic functionality over factual information (*e.g.*, drug-disease targeting prediction and distinctive disease-gene association identification) is highly desirable.

Proposed Approach. This paper presents a novel system, called Life-iNet, which transforms an *unstructured* corpus into a *structured* network of factual knowledge, and supports multiple exploratory and analytic functions over the constructed network for knowledge discovery. Life-iNet automatically detects token spans of entities mentioned from text, labels entity mentions with semantic categories, and identifies relationships of various relation types between the detected entities. These inter-related pieces of information are integrated to form a unified, structured network, where nodes represent different types of entities and edges denote relationships of different relation types between the entities (see Fig. 1 for example). To address the issue of limited diversity and coverage, Life-iNet relies on the external knowledge bases to provide seed examples (*i.e.*, *distant supervision*), and identifies additional entities and relationships from the given corpus (*e.g.*, using multiple text resources such as scientific literature and encyclopedia articles) to construct a structured network. By doing so, we integrate the factual information in the existing knowledge bases with those extracted from the corpus. To support analytic functionality, Life-iNet implements link prediction functions over the constructed network and integrates a distinctive summarization function to provide insight analysis (*e.g.*, answering questions such as "*which genes are distinctively related to the given disease type under* `GeneDiseaseAssociation` *relation?*").

To systematically incorporate these ideas, Life-iNet leverages the novel distantly-supervised information extraction techniques (Ren et al., 2017, 2016a, 2015) to implement an *effort-light network construction framework* (see Fig. 2). Specially, it relies on distant supervision in conjunction with external knowledge bases to (1) detect quality entity mentions (Ren et al., 2015), (2) label entity mentions with fine-grained entity types in a given type hierarchy (Ren et al., 2016a), and (3) identify relationships of different types between entities (Ren et al., 2017). In particular, we design specialized loss functions to faithfully model "*appropriate*" labels and remove "*false positive*" la-

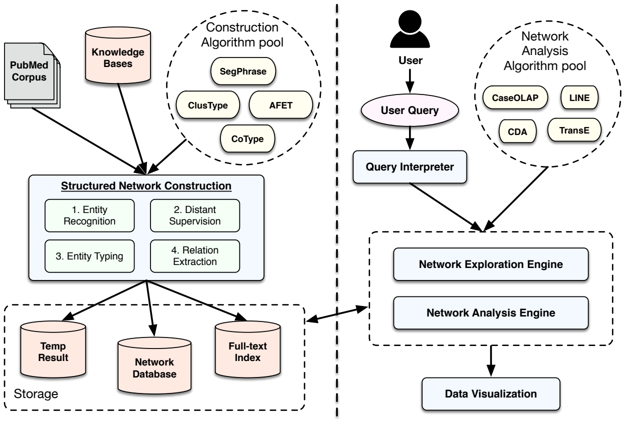


Figure 2: System Architecture of Life-iNet.

bels for the training instances (heuristically generated by distant supervision), regarding the specific context where an instance is mentioned (Ren et al., 2017, 2016a). By doing so, we can construct *corpus-specific* information extraction models by using distant supervision in a noise-robust way. The proposed network construction framework is domain-independent—it can be quickly ported to other disciplines and sciences without additional human labeling effort. With the constructed network, Life-iNet further applies link prediction algorithms (Tang et al., 2015; Bordes et al., 2013) to infer new entity relationships, and distinctive summarization algorithm (Tao et al., 2016) to find other entities that are distinctively related to the query entity (or the given entity types).

Contributions. The contributions and features of the Life-iNet system are summarized as follows.

1. A novel knowledge exploration and analysis system for life sciences that integrates existing knowledge bases and factual information extracted from massive literature.
2. An *effort-light* framework that leverages distant supervision in a *robust* way to automatically construct a structured network of factual knowledge from the given unstructured text corpus.
3. Capabilities for exploration and analysis over the constructed structured network to facilitate life sciences research.

The Life-iNet demo system will be made available online for interactive use after its demonstration in the conference.

2 The Life-iNet System

At a high level, Life-iNet consists of two major components: a structured network construction pipeline and a network exploration and analysis engine. The former (*i.e.*, the network construction pipeline) includes four functional modules: (1) entity mention detection, (2) distant supervision generation, (3) entity typing, and (4) relation extraction; whereas the latter (*i.e.*, the network exploration and analysis engine) implements network exploratory functions, relationship prediction algorithms (*e.g.*, LINE (Tang et al., 2015)) and network-based distinctive summarization algorithms (*e.g.*, CaseOLAP (Tao et al., 2016)), and operates on the constructed network to support answering different user queries. Fig. 2 shows its system architecture. The functional modules are presented in detail as follows.

Background corpora	Cancer	Heart Disease
#PubMed publications	2,936,615	2,105,257
#PMC full-text papers	95,008	38,205
#Wikipedia articles	37,128	25,577
#Sentences in total	38M	23M
#Entity types	1,116	1,086
#Relation types	414	384
#KB-mapped (seed) entity mentions	59M	33M
#KB-mapped (seed) relation mentions	47M	23M
#Nodes in Life-iNet (*i.e.*, entities)	64M	39M
#Edges in Life-iNet (*i.e.*, facts)	186M	82M

Table 1: Data statistics of corpora and networks in Life-iNet.

2.1 Structured Network Construction

The network construction pipeline automatically extracts factual structures (*i.e.*, entities, relations) from given corpora with (potentially noisy) distant supervision, and integrates them with existing knowledge bases to build a unified structured network. In particular, to extract high-quality, typed entities and relations, we design *noise-robust objective functions* to select the "*most appropriate*" training labels when constructing models from labeled data (heuristically obtained by distant supervision) (Ren et al., 2016b,a, 2017).

Data Collection. To obtain background text corpora for network construction, we consider two kinds of textual resources, *i.e.*, scientific publications and encyclopedia articles. For scientific publications, we collect titles and abstracts of 26M papers from the entire PubMed[1] dump, and full-text paper content of 2.2M papers from PubMed Central (PMC)[2]. For encyclopedia articles, we collect 62,705 related articles through Wikipedia Health Portal[3]. For demonstration purpose, we select documents related to two kinds of important diseases, *i.e.*, cancer and heart diseases to form the background corpora for Life-iNet. Table 1 summarizes the statistics of the background corpora.

Entity Mention Detection. The entity mention detection module in Life-iNet runs a data-driven

[1] https://www.ncbi.nlm.nih.gov/pubmed/

[2] https://www.ncbi.nlm.nih.gov/pmc/

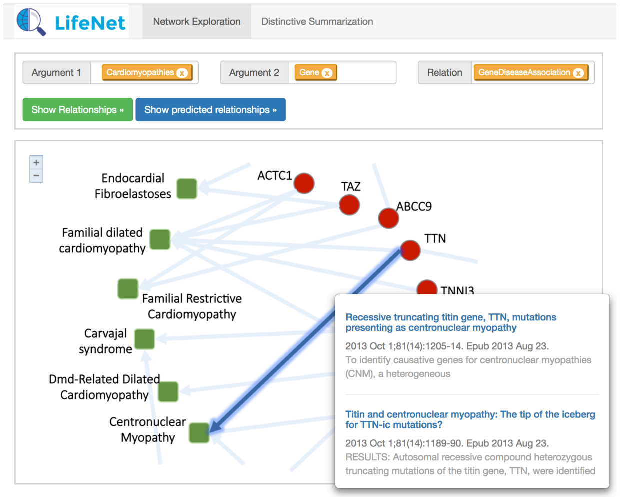
[3] https://en.wikipedia.org/wiki/Portal:Health_and_fitness

text segmentation algorithm, SegPhrase (Liu et al., 2015), to extract high-quality words/phrases as entity candidates. SegPhrase uses entity names from KBs as positive examples to train a quality classifier, and then efficiently segments the corpus by maximizing the joint probability based on the trained classifier. Table 1 shows the statistics of detected entity mentions for the corpora.

Distant Supervision Generation.. Distant supervisions (Mintz et al., 2009; Ren et al., 2017, 2016a) leverages the information overlap between external KBs and given corpora to automatically generate large amounts of training data. A typical workflow is as follows: (1) map detected entity mentions to entities in KB, (2) assign, to the entity type set of each entity mention, KB types of its KB-mapped entity, and (3) assign, to the relation type set of each entity mention pair, KB relations between their KB-mapped entities. Such a label generation process may introduce noisy type labels (Ren et al., 2017). Our network construction pipeline faithfully incorporates the noisy labels in training to learn effective extraction models. In Life-iNet, we use a publicly-available KB, UMLS (Unified Medical Language System)[4], and further enrich its entity type ontology with MeSH tree structures[5]. This yields a KB with 6.7M unique entities, 10M entity relationships, 56k entity types, and 581 relation types. Table 1 shows the data statistics of distant supervision.

Entity Typing. The entity typing module is concerned with predicting a single type-path in the given entity type hierarchy for each *unlinkable* entity mention (*i.e.*, mentions that cannot be mapped to entities in KB) based on its local context (*e.g.*, sentence). Life-iNet adopts a two-step entity typing process, which first identifies the coarse type label for each mention (*e.g.*, `disease`, `gene`, `protein`, `drug`, `symptom`), then refines the coarse label into a more fine-grained type-path (*e.g.*, `disease::heart_disease::arrhythmias`). Specifically, we first run ClusType (Ren et al., 2015) to predict coarse type label for each unlinkable mention. Then, using coarse type label as constraints, we apply AFET (Ren et al., 2016a) to estimate a single type path for each mention. AFET models the noisy candidate type set generated by distant supervision to learn a predictive typing model for unseen entity mentions.

Relation Extraction. The task of relation extraction focuses on determining whether a relationship of interest (*i.e.*, in given relation type set) is expressed between a pair of entity mentions in a specific sentence, and label them with the appropriate relation type if a specific relation is expressed. Life-iNet relies on a distantly-supervised relation extraction framework, CoType (Ren et al., 2017), to extract typed relation mentions from text. CoType leverages a variety of text features extracted from the local context of a pair of entity mentions, and jointly embeds relation mentions, text features and relation type labels into a low-dimensional space, where, in that space, objects with similar type semantics are also close to each other. It then performs nearest neighbor search to estimate the relation type for a relation mention.



Recessive truncating titin gene, TTN, mutations presenting as centronuclear myopathy

2013 Oct 1;81(14):1205-14. Epub 2013 Aug 23.

To identify causative genes for centronuclear myopathies (CNM), a heterogeneous

Titin and centronuclear myopathy: The tip of the iceberg for TTN-ic mutations?

2013 Oct 1;81(14):1189-90. Epub 2013 Aug 23.

RESULTS: Autosomal recessive compound heterozygous truncating mutations of the titin gene, TTN, were identified

Figure 3: Screen shot of the user interface for relation-based exploration and relationship prediction in Life-iNet.

Performance of Network Construction. Performance comparisons with state-of-the-art (distantly-supervised) information extraction systems demonstrate the effectiveness of the proposed pipeline (Ren et al., 2017)—CoType achieves a 25% F1 score improvement on relation extraction and a 6% enhancement in F1 score for entity recognition and typing, on the public BioInfer corpus (manually labeled biomedical papers). Table 1 summarizes the statistics of the constructed structures networks—Life-iNet discovers over 250% more facts compared to those generated by distant supervision.

2.2 Network Exploration and Analysis

The network exploration and analysis engine indexes the network structures and their related textual evidence to support fast exploration. It also implements several network mining algorithms to facilitate knowledge discovery.

[4] `https://www.nlm.nih.gov/research/umls/`

[5] `https://www.nlm.nih.gov/mesh/intro_trees.html`

Network Exploration. For each entity e_i, we index its entity types $\mathcal{T}_i$, and sentences $\mathcal{S}_i$ (and documents $\mathcal{D}_i$) where it is mentioned. For each relation mention $z_i = (e_1, e_2; s)$, we index its sentence s and relation type r_i. With this data model, Life-iNet can support several structured search queries: (1) find entities of a given entity type, (2) find entities that have a specific relation to a given entity (entity type), and (3) find papers related to given entities, entity types, relationships, or relation types. We use raw frequency discounted by object popularity to rank the results.

Relationship Prediction. We adopt state-of-the-art heterogeneous network-based link prediction algorithms, LINE (Tang et al., 2015) and TransE (Bordes et al., 2013), to discover new relationships in the network. The intuition behind these algorithms is straightforward: if two nodes share similar neighbors in the network, they should be related. Following this idea, the algorithms embed the network into a low-dimensional space based on distributional assumption. A new edge will be formed if the similarity between the embedding vectors of the corresponding entity arguments are larger than a pre-defined threshold δ, *i.e.*, $\mathrm{sim}(\mathrm{vec}(e1), \mathrm{vec}(e2)) > \delta$. The prediction can be further interpreted using existing network structures, by retrieving indirect paths between the two entities (if there exists).

Distinctive Summarization. In biomedical domain, some high-popularity entities may form relationships with many other entities simultaneously. For example, some genes may be associated with multiple heart disease types. It is desirable to find genes that are distinctively associated with *each* heart disease type. This motivates us to apply CaseOLAP (Tao et al., 2016), a context-aware, multi-dimensional summarization algorithm to generate distinctive entities. The basic idea is that: an entity is *distinctively* related to the target entity type if it is relevant to entities of the *target* entity type but relatively irrelevant to entities of the *other* entity types. We pre-compute the distinctive summarization results between different entity types and materialize the temporary results for efficient user query answering.

3 Demo Scenarios

3.1 Relation-Based Exploration

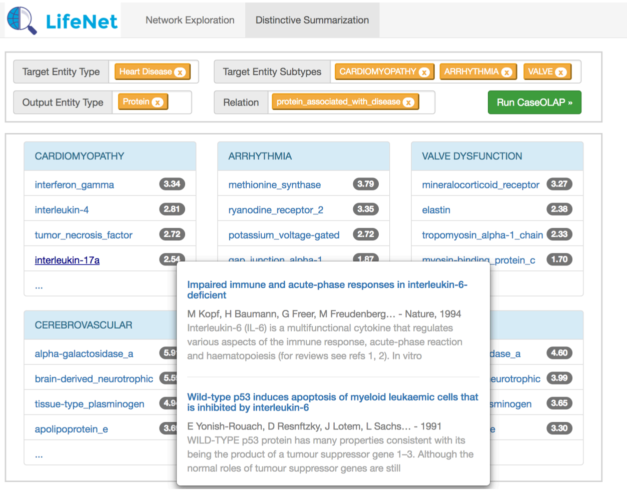
Life-iNet indexes the extracted factual structures along with their support documents. Our demo provides an exploration interface (see Fig. 3), where users can enter an argument triple to specify the entity and relation types they want to explore (user will be prompted with type candidates). Suppose a biologist is interested in finding genes associated with `cardiomyopathies`, he/she can enter type `gene` as argument 1, `cardiomyopathies` as argument 2, and `GeneDiseaseAssociation` as the relation. Life-iNet will then retrieve and visualize a sub-network to show different `cardiomyopathies` entities (*e.g.*, *Endocardial Fibroelastoses*, *Centronuclear Myopathy*, *Carvajal syndrome*), and their associated `gene` entities (*e.g.*, *TAZ*, *BIN1*, *DSC2*). When a user moves his/her mouse cursor to an edge (or node) in the sub-network, Life-iNet will return a ranked list of supporting papers (also linked to PubMed) related to the target relationship (or entity), based on the pre-computed relevance measures. Note that Life-iNet also supports specific entities as input for arguments 1 and 2 in the interface.

3.2 Hypothetical Relationship Generation

In life sciences, some entity relationships (*e.g.*, of type `DrugTargetGene`, `GeneDiseaseAssociation`) may not be explicitly expressed in the existing literature. However, indirect connections between two isolated entities in the constructed network may provide good hints on predicting whether a specific relation exists between them. Life-iNet generates high-confidence predictions of new edges for the constructed network and forms hypothetical entity relationships to facilitate scientific research (*e.g.*, discovering a new drug that can target a specific gene). We integrate this analysis function into our relation-exploration interface. For example, when exploring the sub-network for gene-heart disease associations, users can click on the "*Show Predicted Relationships*" to see hypothetical relationships that Life-iNet generates (highlighted as dash-line edges in the network). In particular, Life-iNet provides explanation of the prediction, using the existing network structures—the indirect paths between two isolated entities will be highlighted when a user clicks on the predicted edge. Thus, a user can further retrieve papers related to the edges on the indirect paths to gain better understanding about the hypothetical relationships.

3.3 Distinctive Entity Summarization

Life-iNet provides a separate user interface for distinctive summarization function (see Fig. 4). In many cases, a user would need to



Target Entity Type: Heart Disease x

Target Entity Subtypes: CARDIOMYOPATHY x ARRHYTHMIA x VALVE x

Output Entity Type: Protein x

Relation: protein_associated_with_disease x

CARDIOMYOPATHY	
interferon_gamma	3.34
interleukin-4	2.81
tumor_necrosis_factor	2.72
interleukin-17a	2.54
...	

ARRHYTHMIA	
methionine_synthase	3.79
ryanodine_receptor_2	3.35
potassium_voltage-gated	2.72
gap_junction_alpha-1	1.87

VALVE DYSFUNCTION	
mineralocorticoid_receptor	3.27
elastin	2.38
tropomyosin_alpha-1_chain	2.33
myosin-binding_protein_c	1.70

CEREBROVASCULAR	
alpha-galactosidase_a	5.9
brain-derived_neurotrophic	5.5
tissue-type_plasminogen	4.9
apolipoprotein_e	3.6
...	

dase_a	4.60
eurotrophic	3.99
sminogen	3.65
e	3.30

Impaired immune and acute-phase responses in interleukin-6-deficient

M Kopf, H Baumann, G Freer, M Freudenberg... - Nature, 1994
Interleukin-6 (IL-6) is a multifunctional cytokine that regulates various aspects of the immune response, acute-phase reaction and haematopoiesis (for reviews see refs 1, 2). In vitro

Wild-type p53 induces apoptosis of myeloid leukaemic cells that is inhibited by interleukin-6

E Yonish-Rouach, D Resnftzky, J Lotem, L Sachs... - 1991
WILD-TYPE p53 protein has many properties consistent with its being the product of a tumour suppressor gene 1–3. Although the normal roles of tumour suppressor genes are still

Figure 4: Screen shot for distinctive summarization function.

compare sets of entities (*e.g.*, proteins) related to several entity types (*e.g.*, different types of heart diseases), to discover the distinctive entities related to each entity type. For example, she may want to know what genes are often associated with `arrhythmia` but are unlikely associated with other kinds of heart diseases such as `cardiomyopathy` and `heart_valve_disease`. Life-iNet allows a user to enter: (1) an entity type to specify the target domain (*e.g.*, `heart_disease`), (2) several sub-types of the target entity type for comparison (*e.g.*, `cardiomyopathy`, `arrhythmia`, `heart_valve_disease`), (3) an entity type to specify the list of related entities (*e.g.*, `protein`), and (4) a relation type (*e.g.*, `protein_associated_with_disease`). With user input queries, Life-iNet produces a structured table to summarize the distinctive entities for each entity sub-type. It also shows the distinctiveness score for each entity. A user can click on each distinctive entity to find documents related to the relationship (similar to the use case in relation-based exploration). An example output of the distinctive summarization for `heart_disease` is shown in Fig. 4.

Acknowledgement

Research was sponsored in part by the U.S. Army Research Lab. under Cooperative Agreement No. W911NF-09-2-0053 (NSCTA), National Science Foundation IIS-1320617 and IIS 16-18481, and grant 1U54GM114838 awarded by NIGMS through funds provided by the trans-NIH Big Data to Knowledge (BD2K) initiative. The views and conclusions contained in this paper are those of the authors and should not be interpreted as representing any funding agencies.

References

Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, and Oksana Yakhnenko. 2013. Translating embeddings for modeling multi-relational data. In *NIPS*.

Patrick Ernst, Amy Siu, Dragan Milchevski, Johannes Hoffart, and Gerhard Weikum. 2016. Deeplife: An entity-aware search, analytics and exploration platform for health and life sciences. In *ACL*.

Jung-jae Kim, Piotr Pezik, and Dietrich Rebholz-Schuhmann. 2008. Medevi: retrieving textual evidence of relations between biomedical concepts from medline. *Bioinformatics* .

Jialu Liu, Jingbo Shang, Chi Wang, Xiang Ren, and Jiawei Han. 2015. Mining quality phrases from massive text corpora. In *SIGMOD*.

Ryan McDonald, Fernando Pereira, Seth Kulick, Scott Winters, Yang Jin, and Pete White. 2005. Simple algorithms for complex relation extraction with applications to biomedical ie. In *ACL*.

Mike Mintz, Steven Bills, Rion Snow, and Dan Jurafsky. 2009. Distant supervision for relation extraction without labeled data. In *ACL*.

Hoifung Poon, Chris Quirk, Charlie DeZiel, and David Heckerman. 2014. Literome: Pubmed-scale genomic knowledge base in the cloud. *Bioinformatics* .

Xiang Ren, Ahmed El-Kishky, Chi Wang, Fangbo Tao, Clare R Voss, and Jiawei Han. 2015. ClusType: effective entity recognition and typing by relation phrase-based clustering. In *KDD*.

Xiang Ren, Wenqi He, Meng Qu, Lifu Huang, Heng Ji, and Jiawei Han. 2016a. AFET: Automatic fine-grained entity typing by hierarchical partial-label embedding. In *EMNLP*.

Xiang Ren, Wenqi He, Meng Qu, Clare R Voss, Heng Ji, and Jiawei Han. 2016b. Label noise reduction in entity typing by heterogeneous partial-label embedding. In *KDD*.

Xiang Ren, Zeqiu Wu, Meng Qu, Clare R. Voss, Heng Ji, Tarek F. Abdelzaher, and Jiawei Han. 2017. CoType: Joint extraction of typed entities and relations with knowledge bases. In *WWW*.

Sebastian Riedel and Andrew McCallum. 2011. Fast and robust joint models for biomedical event extraction. In *EMNLP*.

Damian Szklarczyk, Andrea Franceschini, Stefan Wyder, Kristoffer Forslund, Davide Heller, Milan Simonovic, Alexander Roth, Alberto Santos, Kalliopi P Tsafou, et al. 2014. String v10: protein–protein interaction networks, integrated over the tree of life. *Nucleic acids research* .

Jian Tang, Meng Qu, Mingzhe Wang, Ming Zhang, Jun Yan, and Qiaozhu Mei. 2015. Line: Large-scale information network embedding. In *WWW*.

Fangbo Tao, George Brova, Jiawei Han, Heng Ji, Chi Wang, Brandon Norick, Ahmed El-Kishky, Jialu Liu, Xiang Ren, and Yizhou Sun. 2014. Newsnetexplorer: automatic construction and exploration of news information networks. In *SIGMOD*.

Fangbo Tao, Honglei Zhuang, Chi Wang Yu, Qi Wang, Taylor Cassidy, Lance Kaplan, Clare Voss, and Jiawei Han. 2016. Multi-dimensional, phrase-based summarization in text cubes. *Data Engineering* page 74.

Philippe Thomas, Johannes Starlinger, Alexander Vowinkel, Sebastian Arzt, and Ulf Leser. 2012. Geneview: a comprehensive semantic search engine for pubmed. *Nucleic acids research* 40(1):585–591.