Data Min Knowl Disc (2015) 29:976–998
DOI 10.1007/s10618-014-0396-4

Mining strong relevance between heterogeneous entities from unstructured biomedical data

Ming Ji · Qi He · Jiawei Han · Scott Spangler

Received: 2 March 2014 / Accepted: 17 November 2014 / Published online: 5 February 2015

Abstract Huge volumes of biomedical text data discussing about different biomedical entities are being generated every day. Hidden in those unstructured data are the strong relevance relationships between those entities, which are critical for many interesting applications including building knowledge bases for the biomedical domain and semantic search among biomedical entities. In this paper, we study the problem of discovering strong relevance between heterogeneous typed biomedical entities from massive biomedical text data. We first build an entity correlation graph from data, in which the collection of paths linking two heterogeneous entities offer rich semantic contexts for their relationships, especially those paths following the patterns of top-k selected meta paths inferred from data. Guided by such meta paths, we design a novel relevance measure to compute the strong relevance between two heterogeneous entities, named EntityRel. Our intuition is, two entities of heterogeneous types are

Responsible editor: Fei Wang, Gregor Stiglic, Ian Davidson, Zoran Obradovic.

This work was done when the first author was doing an internship at IBM Almaden Research Center.

M. Ji (✉) · J. Han
University of Illinois at Urbana-Champaign, 201 N. Goodwin Avenue, Urbana, IL, USA
e-mail: mingji1@illinois.edu

J. Han
e-mail: hanj@illinois.edu

Q. He
LinkedIn Inc., Mountain View, CA, USA
e-mail: qhe@linkedin.com

S. Spangler
IBM Almaden Research Center, 650 Harry Road, San Jose, CA, USA
e-mail: spangles@us.ibm.com

strongly relevant if they have strong direct links or they are linked closely to other strongly relevant heterogeneous entities along paths following the selected patterns. We provide experimental results on mining strong relevance between drugs and diseases. More than 20 millions of MEDLINE abstracts and 5 types of biological entities (Drug, Disease, Compound, Target, MeSH) are used to construct the entity correlation graph. A prototype of drug search engine for disease queries is implemented. Extensive comparisons are made against multiple state-of-the-arts in the examples of Drug–Disease relevance discovery.

Keywords Biomedical text data · Heterogeneous · Meta path · Relevance · Context-aware

1 Introduction

Recently, data sets containing heterogeneous entities interacting with each other have been found to be increasingly prevalent in real life applications. Examples include drugs, diseases, compounds, targets and so on in the biomedical area, users, items and tags in recommendation systems, authors, papers, venues and keywords in the bibliographic data, etc. Two entities are called heterogeneous typed entities if they are of different types (e.g., drug and disease, user and tag), and therefore have different semantic meanings and properties. Discovering strong relevance relationships between heterogeneous typed entities is a fundamental problem. By strong relevance we mean the relevance supported by rich relevance contexts in the data. Given an entity (e.g., a disease) as a query, a user may be interested in browsing other entities of heterogeneous types (e.g., drugs) that have strong relevance relationships with the queried entity. Similarly, in an online recommendation system, given an item (e.g., a document), it will be very helpful if we can find a set of tags that are strongly relevant to the item and present them to the user. With the discovery of strong semantic relationships between entities, huge knowledge networks can be built, and the user can navigate from one entity to other related entities and quickly find the information he/she is searching for.

Data may come from many different sources. In this paper we focus on unstructured biomedical data for a couple of reasons: (1) For structured data, it may not be so challenging because basic relationships, such as customers "purchase" items, are explicit already, which could be used to derive strong relevance (Jeh and Widom 2002, 2003; Sun et al. 2011; Shi et al. 2012; Lao and Cohen 2010). (2) It is much more challenging to find such relationships in text data, which is unstructured, noisy, and entity relationships are deeply hidden. Moreover, text data are ubiquitous, in huge amount, and being updated constantly. Mining entity relationships from text data is thus imperative. (3) The rich biomedical domain knowledge offers the feasibility of extracting entities from unstructured biomedical data. However, entity recognition from many other domains is still an open problem. (4) Structured data in the biomedical domain is not as widely available as other areas such as social networks, recommendation systems and so on. Instead, research papers, patents and news articles in the form of unstructured text data are more prevalent in the biomedical area. And those unstructured text data are easily accessible. Therefore, mining unstructured data is a critical problem in biomedical knowledge discovery.

In the biomedical domain, drug discovery studies (Searls 2005; Gunther et al. 2003; Coulet et al. 2011; Ramakrishnan et al. 2008) can only detect drugs that are known to treat certain diseases, and cannot discover strong relevance between drugs and diseases that are not explicitly written in the text. In this paper, we try to infer the relevance relationships between biomedical entities based on the semantic information encoded in the text. For example, drug "tretinoin" and disease "acne" are strongly relevant since tretinoin can be used to treat acne, and this strong relevance can potentially be discovered from an unstructured text data corpus in which none of the articles explicitly says that tretinoin treats acne. Recommendation systems suggest the items that the users are likely to be interested in Sen et al. (2009), Yin et al. (2010), Guan et al. (2010). However, the data there are usually structured and the systems require the availability of training data (e.g., some users are interested in certain items). Recent studies on similarity search in heterogeneous graphs, such as PathSim Sun et al. (2011), explore a meta path based similarity measure. Nevertheless, their similarity measure is defined for comparing nodes of the same types (*e.g.*, similarity between authors in a bibliographic network). In other words, (Sun et al. 2011) cannot be applied to discovering relevance between different types of entities unless we ignore all the type information and treat different types of entities as the same type, which, however, violates the philosophy of Sun et al. (2011) which emphasizes the type information on the nodes in heterogeneous graphs. Shi et al. (2012) first proposed to study the relevance between heterogeneous entities. However, their similarity measure is based on pairwise random walk which may not be able to capture the subtlety of the path-constrained strong relevance relationships as indicated in our experiments.

Based on these considerations, we propose our approach, which contributes to the state-of-the-art in the following aspects: (1) the method constructs a biomedical entity correlation graph from unstructured data, extending the scope of the study to unstructured text data; (2) the method extends the meta path based relationship analysis (Sun et al. 2011)from mining relevance between homogeneous typed entities to heterogeneous typed ones and infers top-k most effective meta paths from data; (3) our new approach, EntityRel, proposes a new measure for computing the context-aware relevance between two heterogeneous entities; and (4) our experiments and performance comparison with several existing methods demonstrate the effectiveness of our method, with interesting results in the biomedical domain for the strong relevance discovery between drugs and diseases.

The biomedical entity correlation graph maintains basic entity relationships that can be straightforwardly found in unstructured text data, i.e., weighted co-occurrence. Based on it, the collection of paths linking two heterogeneous entities e_i and e_j offer rich semantic contexts for their relationships. However, not all paths carry the same semantics. For example, "tretinoin–skin–acne" indicates a therapeutic relationship between drug "tretinoin" and disease "acne", while "vitamin A–toxicity–acne" indicates a side-effect relationship. Therefore, the relevance type depends on the contexts in paths. Our proposed approach, EntityRel, is such a context-aware relevance measure. Without loss of generality, we predefine 5 types of biological entities for constructing the entity correlation graph, which are: Drug, Compound, Disease, Target and MeSH. Based on them, we can define path types like "Drug–Target–Disease" or "Drug–MeSH–Disease", named as meta paths in the paper. For example, "tretinoin–

skin–acne" is one path instance of meta path "Drug–Target–Disease". In our approach, EntityRel, we assume that the relevance is only meaningful under path contexts constrained by certain meta path. For example, if we use all paths following the pattern "Drug–Target–Disease" as contexts, the discovered relationships between drugs and diseases are very likely therapeutic relationships. More specifically, we name the set of entities (excluding e_i and e_j) in these paths as "reasoning entities", which are used to reason the discovered relevance relationships.

Consequently, one natural question is: what kinds of paths should we use for mining the strong relevance between heterogeneous entities? The definition of "strong" relevance is a data dependent concept: some types of relevance might be strong and some types might be weak, depending on how rich the corresponding relevance contexts provided by the data can be. In this paper, given two types of entities, we automatically pick up top-k meta paths from the data, such that the relevance contexts defined by these meta paths in data are relatively richer than other types of contexts. Based on these rich contexts, we are supposed to discover so-called "strong" relevance between the given two types of entities.

The remaining of the paper is organized as follows. Section 2 summarizes the related work. We formally define the problem and propose the framework of our solution in Sect. 3. As the first step, in Sect. 4, we build the biomedical entity correlation graph from unstructured data, upon which representative meta paths are generated in Sect. 5. The new measure, EntityRel, is developed upon the entity correlation graph and selected meta paths for the final goal in Sect. 6. Section 7 reports our experimental results and a prototype drug search engine built based on the proposed framework. Finally, Sect. 8 concludes the paper.

2 Related work

As pointed out in Sheth et al. (2005), Anyanwu et al. (2005), the relationships between entities are the heart of the Semantic Web. Substantial efforts are made to develop techniques for searching complex relationships between entities (Anyanwu et al. 2005; Aleman-Meza et al. 2003; Anyanwu and Sheth 2003). The relationships are often referred to as Semantic Associations. However, those Semantic Associations studied in Semantic Web are mainly based on the RDF model, therefore are restricted to simple, existing relationships, such as the "purchase" relationship between customers and items, and the "work for" relationship between professors and universities. Different from such existing work, we focus on discovering meaningful relationships that do not exist in any structured data, but could be inferred from the massive text data.

In the biomedical domain, it is recognized that the text data describing different types of biological entities could be employed to facilitate drug discovery (Searls 2005). Gunther et al. (2003) performs classification over the drug-induced genomic expression profiles to predict the clinical drug efficacy. Different from them, we hope to discover strong relevance in an unsupervised way using the general text corpus. Natural language processing techniques have also been adopted to mine relationships between biological entities from the text data (Coulet et al. 2011; Ramakrishnan et al. 2008). However, similar to the Semantic Web technologies, the approaches based on

natural language processing can only detect relationships that are already expressed by words or phrases in text. On the contrary, we focus on discovering strong relevance between drugs and diseases that may not necessarily have been written in the text, which is much more useful for new drug discovery.

Another family of related work is the recommendation systems, which suggest the items that users are likely to be interested in Sen et al. (2009), Yin et al. (2010), Guan et al. (2010). Although recommendation also discovers unknown relationships, our problem is fundamentally different from the classical recommendation problem. First, we aim to develop a fully automatic approach that does not use any label information, while recommendation systems usually know some users are interested in certain items. Second, the data source we hope to discover strong relevance from is the text corpus, without the attributes and profiles of the entities, or some basic relationships (such as user-tag and item-tag) among entities in typical recommendation scenarios.

Given a graph, many methods have been developed for estimating relevance between two nodes. Personalized PageRank (Jeh and Widom 2003) and SimRank (Jeh and Widom 2002) are two representatives for computing the similarity between two nodes of the same type in a homogeneous graph. However, in our problem, different types of nodes carry different semantic meanings and should not be mixed (Sun et al. 2011). For heterogeneous graphs, PathSim (Sun et al. 2011) gives an interesting meta path based similarity measure between two nodes of the same type. HeteSim (Shi et al. 2012) and Path Constrained Random Walk (Lao and Cohen 2010) estimate the relevance between different types of nodes following the random walk framework. However, the original HeteSim algorithm only uses the binary graph, ignoring the weight on the edges, which is shown to be critically important in our experiments. Path Constrained Random Walk incorrectly favors the popular entities and ignores the differences of various contexts inherited from various meta paths. More discussion about these methods can be found in Sect. 6.1.

3 Problem and framework

Given an unstructured biomedical text data corpus $\mathcal{D}$ and K types of predefined biological entities $E_1, \ldots, E_K$, our problem is to automatically discover the strong relevance relationships between any pair of entities e_i and e_j strongly supported by $\mathcal{D}$, where e_i and e_j can belong to either the same entity type or different entity types. As a more general case, in this paper we focus on the relevance relationships across heterogeneous entity types. We annotate $E(e_i)$ as the entity type name of e_i and $|E(e_i)|$ as the number of entities of type $E(e_i)$.

Formally, we quantify the relevance relationship between two heterogeneous entities e_i and e_j as a relevance score $R(e_i, e_j)$.

The computation of $R(e_i, e_j)$ depends on the observed correlations of e_i and e_j in data $\mathcal{D}$. Possibly the simplest correlation between e_i and e_j is the number of co-occurrence in $\mathcal{D}$. However, the simple co-occurrence model can not effectively capture the correlation contexts of e_i and e_j. For example, given a frequent sentence "Which one is able to treat acne, doxycycline or tetracycline?" in MEDLINE, it is hard to tell the drug entity "tetracycline" is relevant to the disease entity "acne" or not.

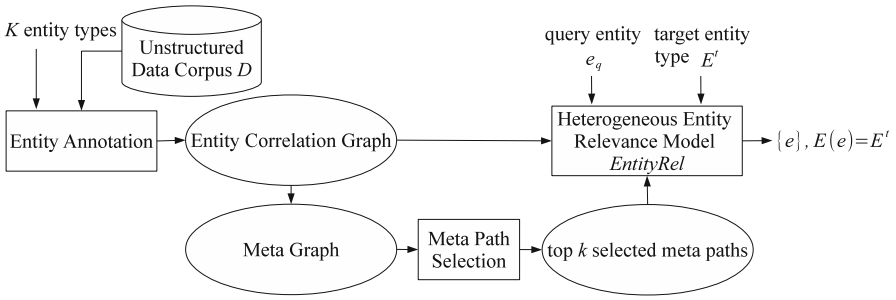


Fig. 1 System framework of EntityRel

We observe that, the correlation contexts between two entities can be manifested by other entities that frequently co-occur with both. For example, given the frequent sentence "Acne is a disease that affects the skin's oil glands." in MEDLINE, we know that to treat the disease "acne", the organism entity "skin" is one kind of targets. Then, given another frequent sentence "Tetracyclines are oral antibiotics often used to treat skin diseases." in MEDLINE, we know that the function scope of drug "tetracycline" covers the target entity "skin". Thus, the target entity "skin" effectively links the drug entity "tetracycline" and the disease entity "acne" together and implies their relevance.

The rich context information in unstructured data corpus $\mathcal{D}$ can be represented by an undirected Entity Correlation Graph $\mathcal{G}$. In graph $\mathcal{G}$, the nodes are heterogeneous entities and the edge between two entity nodes represents the fact that these two entities once co-occur in data $\mathcal{D}$. Given one node e_i, its neighborhood set $N(e_i)$ includes all other entities that co-occur with it in data. Given two example entities "tetracycline" and "acne", we can extract a number of paths linking them, e.g., "tetracycline–skin–acne" and, "tetracycline—protein synthesis inhibitor—bacterial infection—acne", from graph $\mathcal{G}$. All these paths collectively serve as the correlation contexts for "tetracycline" and "acne".

Compared to the unstructured data corpus $\mathcal{D}$, the Entity Correlation Graph $\mathcal{G}$ is structured and easy to analyze and store. Representing the unstructured text as such a graph enables the task of entity relationship discovery to be formulated as searching relevant heterogeneous entities by traveling the graph. For example, for discovering the relationships between drugs and diseases, the user inputs a disease entity "acne" and then searches all drug entities reachable in the graph. Without the loss of generality, we formulate $R(e_i, e_j)$ as a search problem: computing $R(e_q, e)$ by treating $e_q = e_i$ as the query entity, $e = e_j$ as the searching target entity, and $E(e) = E(e_j)$ which is the target entity type. The whole framework is given by Fig. 1.

4 Construction of correlation graph

Many existing work in graph-based entity search assume the existence of the entity graph (Jeh and Widom 2002, 2003; Sun et al. 2011; Shi et al. 2012; Lao and Cohen 2010). However, in this paper, how to automatically generate an Entity Correlation

Graph $\mathcal{G}$ from the unstructured data corpus $\mathcal{D}$ remains challenging, which is discussed in this section.

4.1 The unstructured data corpus $\mathcal{D}$

We use MEDLINE,[1] a bibliographic database of life sciences and biomedical information, as the knowledge base to discover entity relationships in the paper. The abstracts of all 20,642,063 biomedical documents to date consist of the unstructured data corpus $\mathcal{D}$.

We select 5 representative types of biological entities, *Drug*, *Disease*, *Compounds*, *Target* and *MeSH terms*, to study in the paper. In total, we predefined 5,867 FDA-approved drugs;[2] a dictionary of 4,244 diseases extracted from human disease ontology;[3] a set of 2,254 small-molecule chemical compounds with explicit drug indications from the Chemical Entities of Biological Interest (ChEBI) database;[4] a dictionary of 11,280 targets made up of 4 sub types: tissue, cell-line, protein, and organism; and a set of all 17,347 leaf MeSH terms in the MeSH tree,[5] which are used as the meta-data to index medical articles in MEDLINE by NIH. All the above entities consist of the node set in the Entity Correlation Graph G.

4.2 Entity annotation in text

Given the MEDLINE corpus and 5 types of biological entities, the first step is to annotate those entities in the MEDLINE articles.

For Disease and Target annotators, we adopt a dictionary-based method to look up entities in the text based on the exact string match.

For Drug annotator, considering a drug usually contains a number of synonyms like brand name etc., our method is dictionary-based and enhanced by synonyms extracted from CHEMBL.[6]

For Compound annotator, we designed a context-aware Conditional Random Field model (Yan et al. 2011), where both compound structural features and text content features are used to infer the labeling of text. To reduce the ambiguity of compound names, we convert all compound substances to their International Chemical textual identifiers (InChI) first, and then used InChIKey, a fixed length (25 character) condensed digital representation of InChI, to represent each compound.

For MeSH annotator, as all articles in MEDLINE already have 10–15 MeSH terms labeled by human, we simply used these labeled MeSH terms as the annotation results.

[1] http://www.nlm.nih.gov/bsd/pmresources.html.

[2] http://www.accessdata.fda.gov/scripts/cder/drugsatfda/.

[3] http://www.obofoundry.org/cgi-bin/detail.cgi?id=disease_ontology.

[4] http://www.ebi.ac.uk/chebi/. Note that drugs belong to compounds. In this paper, we treat them differently as they originate from different sources orthogonally.

[5] http://www.nlm.nih.gov/mesh/.

[6] https://www.ebi.ac.uk/chembl/.

4.3 Correlation weight in correlation graph

After entities are annotated in text, we can easily add an edge between two entities e_i and e_j in the Entity Correlation Graph if they are annotated in the same set of articles. The remaining question is to find a reasonable weighting function w_{ij} for the edge. Straightforwardly, we can simply use the raw number of co-occurrence to weigh the edge $w_{ij} = co(e_i, e_j)$, where $co(e_i, e_j)$ is the number of articles where both e_i and e_j are annotated in the text. However, this simple method largely favors those popular entities appearing in many articles. Instead, we propose to compute w_{ij} with full consideration of both the relevant frequency that two entities co-occur and the popularity of each entity. Following the classical TF-IDF model in information retrieval, we assume a large w_{ij} implies:

(1) e_i and e_j co-occur frequently;
(2) e_i (or e_j) occurs rarely with other entities of the type $E(e_j)$ (or $E(e_i)$).

To satisfy the first condition, we design a normalized symmetric frequency function as

$$freq(e_i, e_j) = \frac{co(e_i, e_j)}{\left(\sum_{e_y \in E(e_j)} co(e_i, e_y) + \sum_{e_x \in E(e_i)} co(e_x, e_j)\right)/2}.$$

To satisfy the second condition, we define $ief(e_i, e_j)$ which represents the "inverse entity frequency" to measure whether entities e_i and e_j are common or rare within all the co-occurrence between entities of types $E(e_i)$ and $E(e_j)$:

$$ief(e_i, e_j) = \log \frac{(|E(e_i)| + |E(e_j)|)/2}{1 + \left(|N(e_i) \wedge E(e_j)| + |N(e_j) \wedge E(e_i)|\right)/2}, \tag{1}$$

where $N(e_i) \wedge E(e_j)$ represents the joint set of entities who are neighborhoods of e_i and have the same entity type as e_j.

Finally, we have the symmetric correlation weighting function

$$w_{ij} = freq(e_i, e_j) \times ief(e_i, e_j). \tag{2}$$

Note that this correlation weighting function is different from this paper's target function $R(e_i, e_j)$. The former is designed to weigh the correlation between two entities without considering the global correlation graph structure and other type of entities. It can be treated as one kind of local affinity measure between two entities. This function can be used as one naive solution of $R(e_i, e_j)$. But, this naive solution undoubtedly has many limitations. First, the correlation contexts which have been previously shown to be effective in linking entities are lost. Second, it cannot find those entities which never directly co-occur with the query entity.

In this paper, we use w_{ij} as the elementary edge weighting function for the Entity Correlation Graph G; and then explore other more sophisticated graph travel methods for the entity relationship discovery based on the graph.

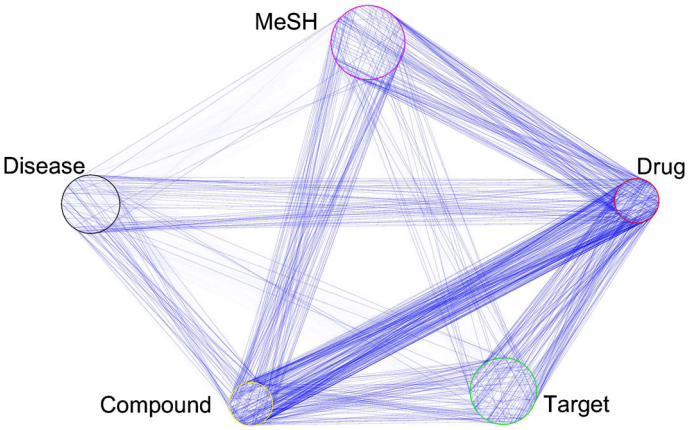


Fig. 2 Entity correlation graph $\mathcal{G}$. *Each edge* represents 1,000 links in data, where the intensity of the color represents the weights of the links. The size of *each circle* is proportional to the number of entities of each type

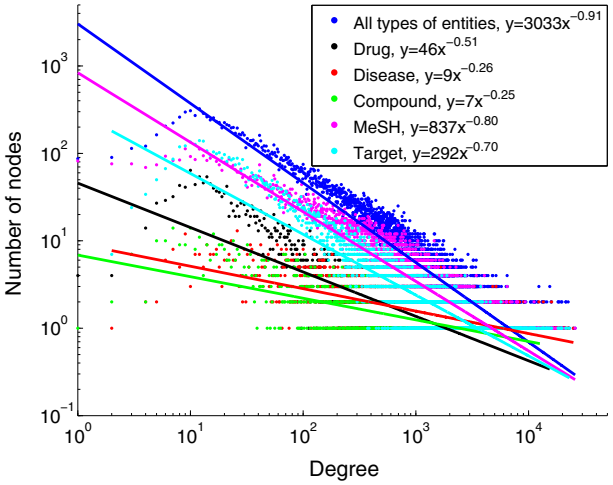


Fig. 3 Degree distribution of the nodes in $\mathcal{G}$

4.4 Properties of entity correlation graph

The constructed Entity Correlation Graph visualized by Fig. 2 has many interesting properties. It tells some meta paths (*e.g.* Drug–Compound) are much denser than others (*e.g.* MeSH–Disease). To uncover strong relevance from the data, we may only focus on those dense meta paths which are strongly supported by data.

The degree distributions of $\mathcal{G}$ of individual entity types are depicted in Fig. 3. One interesting finding is: various entity types have various degree distributions, resulting in various graph structures. For example, both Disease and Compound have very flat power law slope, indicating that their node degrees are more uniformly distributed.

Relatively, the other entity types contain fewer highly connected nodes. This finding discloses that, if we treat the entire graph as a homogeneous graph without differentiating entity types and then randomly surf in graph, some entity types will be favored and some entity types are not reachable. Therefore, traditional methods to compute entity relevance in a homogeneous graph like SimRank (Jeh and Widom 2002) and PathSim (Sun et al. 2011) are not suitable for relevance between heterogeneous entities.

Graph $\mathcal{G}$ is a typical "small world". 91.75 % of its nodes belong to a giant connected component. The average distance between two nodes in this giant component is 2.0663, indicating that starting from one node, we can quickly arrive at other nodes. The "small world" phenomenon in $\mathcal{G}$ offers rich contexts (numerous different paths) between two nodes.

5 Meta path for correlation contexts

With entity correlation graph, the next step is to learn strong correlation contexts from it for discovering strong entity relevance relationships in this section.

5.1 Strong meta paths as contexts

As we mention before, given the correlation graph $\mathcal{G}$, we can formulate the task of entity relevance relationship discovery as searching relevant heterogeneous entities in the graph. For example, given the disease "acne", what are the similar drugs in the graph? The objective of the problem is to infer the relevance score $R(e_q, e)$, given the query entity e_q and one entity e of the target entity type.

In Sect. 4.4, graph $\mathcal{G}$ has been shown to be extremely complicated and overwhelmed across different entity types. Given two heterogeneous entities e_q and e, there exist numerous paths linking them if we do not constrain the length of path and types of entities in the path. More specifically, due to the "small world phenomena" in $\mathcal{G}$, most pairs of entities can be linked together within 2 steps. Apparently, we do not want to recommend all entities to the query. From the complex graph $\mathcal{G}$, we observe that,

(1) If a path is too long, it may cause concept drift and link two irrelevant entities together. For example, a path "acne (Disease)–skin (Target)–muscle (Target)–Ryanodine Receptor Calcium Release Channel (MeSH)–rycals (Drug)" links the drug "rycals" that treats skeletal muscle to an irrelevant disease "acne".
(2) To study the strong relevance relationships between two types of entities, some types of paths are preferred to others for a given data. For example, in Fig. 2, the meta path "Drug–Compound–Disease" has many more path instances than the meta path "Drug–MeSH–Disease". The former thus implies stronger relevance than the latter in data. Or, the relevance implied by the latter is less confident, due to the sparsity of the data.

Our observations indicate that, to compute $R(e_q, e)$, a subset of relatively short paths between e_q and e should be extracted from the graph $\mathcal{G}$ to serve as the correlation contexts. Finding a maximum induced connected subgraph (all paths are connected to e_q and e) with certain property (i.e., following the above two observations) from $\mathcal{G}$ is

a classic NP-complete problem. For that reason, we design an approximate solution for context selection follows.

We first convert the context selection problem into a meta path selection problem. As long as a meta path is selected, all its path instances will be selected as contexts. The meta path is formally defined as follows.

Definition 1 Meta Path. A meta path M of length l is a sequence of nodes in the form of $E_{x_1} \xrightarrow{A_{x_1,x_2}} E_{x_2} \xrightarrow{A_{x_2,x_3}} \cdots \xrightarrow{A_{x_l,x_{l+1}}} E_{x_{l+1}}$ where $x_y \in [1, K]$, $y \in [1, l]$. $A_{x_y,x_{y+1}}$ defines a composite correlation weight between two entity types E_{x_y} and $E_{x_{y+1}}$; and $A_{x_1,x_2} \cdots A_{x_l,x_{l+1}}$ defines a composite correlation weight W_M for the path M:

$$\begin{aligned} W_M &= A_{x_1,x_2} \cdots A_{x_l,x_{l+1}} = \prod_{y=1}^{l} A_{x_y,x_{y+1}} \\ &= \prod_{y=1}^{l} \frac{\sum_{e_i \in E_{x_y}, e_j \in E_{x_{y+1}}} w_{ij}}{|E_{x_y}| \times |E_{x_{y+1}}|}, \end{aligned} \tag{3}$$

where w_{ij} denotes the correlation for regular paths (rf. Eq. 2).

Meta path defines the sequence pattern of regular paths. A meta path with large path weight implies that the regular path instances following its pattern have a large correlation weight on average. Our meta path weight definition favors the short paths. To compute $R(e_q, e)$ by ranking all meta paths starting from $E(e_q)$ and ending at $E(e)$ w.r.t their weights, if we only choose top-k meta paths, the discovered relevance relationships are relatively strong in data.

One may wonder that why our context selection strategy is at the meta path level, not at the path instance level. We have two reasons. First, context selection at the path instance level has to be computed on-line for each query. Due to the large size of entity correlation graph, it is thus not efficient to deploy in the real time. Instead, context selection at the meta path level can be computed off-line. Second, no single data can cover all possible entity relationships in the real world. Thus, the path instance level context selection may overfit the data. Alternatively, context selection at the meta path level strikes a good balance between data sparsity and the average performance.

5.2 Meta graph for meta path selection

Given K types of entities, there are $K(K-1)/2$ different pairs of types. For each pair, we need pre-compute all possible meta paths and rank them w.r.t. their weights. One simple yet efficient way to enumerate the meta paths is to maintain a meta graph in memory, which is defined as follows.

Definition 2 Meta Graph. Given the entity correlation graph $\mathcal{G}$ and K types of entities $E_1, \ldots, E_K$, a graph $\mathcal{G}^m$ is called a meta graph over $\mathcal{G}$ when its nodes are one of K entity types and the weight between two entity types E_i and E_j is defined as $A_{i,j}$ that follows the composite correlation weight definition in Eq. 3.

ALGORITHM 1: Top-k meta paths selection

Input: Meta graph $\mathcal{G}^m$ and two question entity types E_i and E_j for entity relationship discovery; k
Output: Top-k meta paths in terms of path weights.
Initialize two empty sets A and $\mathcal{O}$;
repeat
  Find the meta path M not in $\mathcal{O}$ with the highest path weight from $\mathcal{G}^m$. Path M must have the form of $E_i \xrightarrow{A_{i,x_2}} E_{x_2} \xrightarrow{A_{x_2,x_3}} \cdots \xrightarrow{A_{x_l,j}} E_j$;
  Insert all pairs of composite correlations $A_{x_y,x_{y+1}}$ in path M into the set A;
  if $|A|$ *increases* **then**
    Insert path M into $\mathcal{O}$;
  end
until $|\mathcal{O}| = k$;
return $\mathcal{O}$

The meta graph $\mathcal{G}^m$ actually defines a $K \times K$ pair-wise weight matrix for K entity types. It is a dense graph because $A_{i,j} > 0$ as long as there exists one entity of type E_i that co-occurs with another entity of type E_j in data. It is a symmetric graph as $A_{i,j} = A_{j,i}$. The diagonal elements in the matrix (when $i = j$) indicate the self-correlations of one entity type, which cannot be ignored because it is common that entities of the same type co-occur with each other. The meta graph can be seen as a summary of the original large entity correlation graph at the entity type level. For example, based on our entity correlation graph $\mathcal{G}$ built upon MEDLINE and five types of entities (Disease, Drug, Compound, Target, MeSH), we build a meta graph $\mathcal{G}^m$.

Based on the meta graph and the starting/ending entity types, we can efficiently enumerate all possible meta paths. Recall that we have two principles to select meta paths: (1) the length is not too long; and (2) the path has high weight. As our meta path weighting function (Eq. 3) implicitly favors short paths, a simple greedy algorithm that travels all meta paths from the highest path weight to the lowest path weight is sufficient to our goal, as shown in Algorithm 1. Each time a new meta path M is selected, it must contain at least one new type of composite correlation, or a new pair of entity types in another word. This heuristic prevents the information self loop in meta path and thus largely limits the scope of path candidates.

5.2.1 Time complexity of top-k meta paths selection

The time complexity of Algorithm 1 is $O(K! \log(K!))$. As searching the highest weighted path is the most time consuming part, one trick to further improve the efficiency is to take the logarithm of the meta path weighting function as follows:

$$\begin{aligned} \arg\max_M \log W_M &= \arg\max_M \sum_{y=1}^{l} \log A_{x_y,x_{y+1}} \\ &= \arg\min_M \sum_{y=1}^{l} -\log A_{x_y,x_{y+1}}. \end{aligned}$$

Table 1 Top 5 meta paths selected by Algorithm 1 and their weights

Meta path	Weight
Drug–disease	8.4×10^{-5}
Drug–drug–disease	1.0×10^{-7}
Drug–compound–disease	2.9×10^{-8}
Drug–disease–disease	1.3×10^{-8}
Drug–MeSH–disease	1.5×10^{-9}

In our experimental data set, $A_{i,j} < 1$, therefore $-\log A_{i,j} > 0, \forall i, j \in \{1, \ldots, K\}$. With this conversion, our method is converted to finding the k-shortest paths in Eppstein (1998) with time complexity $O(|(K^2 + K)/2| + |K| \log |K| + k)$ (which is actually $O(K^2 + k)$). Given that the number of entity types K is usually not large in real data, this complexity is acceptable.

5.2.2 Example results of top-k meta paths selection

By setting $k = 5$ and selecting two entity types Drug and Disease as an example, the top meta paths generated by Algorithm 1 from our data are listed in Table 1. These five meta paths collectively generate the correlation contexts for any $\langle drug, disease \rangle$ pair while measuring their strong relevance.

6 Meta path based heterogeneous entity relevance model

For the problem of searching relevant heterogeneous entity e of type E^t in graph $\mathcal{G}$ for a query entity e_q, the previous section selects top-k meta paths as the relevance contexts. The top-k meta paths collectively define a subgraph $\mathcal{G}' \in \mathcal{G}$. Based on it, our core task is to compute $R(e_q, e)$.

6.1 Review related work in computing $R(e_q, e)$

The related work in computing $R(e_q, e)$ can be categorized along two dimensions: context-aware and context-agnostic; homogeneous and heterogeneous.

Personalized PageRank (Jeh and Widom 2003) computes the probability that a random walker starting from e_q can arrive at e in the graph as $R(e_q, e)$, where the teleport only switches to the query entity e_q. As a general-purpose graph similarity measure, Personalized PageRank is a context-agnostic model designed for a homogeneous graph. Its variation, called Path Constrained Random Walk (Lao and Cohen 2010), is extended for heterogeneous graphs. It computes the probability of a random walker starting from e_q can arrive at e through constrained paths in the graph as $R(e_q, e)$. However, these random walk models favor the popular entities undesirably (Sun et al. 2011).

SimRank (Jeh and Widom 2002) is another context-agnostic model designed for the homogeneous graph. It iteratively computes $R(e_q, e)$ as the sum of similarities between their neighbors in the graph. The entity types of their neighbors are ignored. HeteSim (Shi et al. 2012) extended SimRank to the heterogeneous graph. Given a

meta path, it computes the average fraction of information that can diffuse from the middle node of the path to two ends as $R(e_q, e)$. However, HeteSim only depends on the raw counts of paths without fully utilizing the rich contexts of these paths.

6.2 Context-aware relevance model

Our goal here is to design a novel relevance measure between two heterogeneous entities that fully considers the subtlety of different types among entities and the top selected meta paths as the correlation contexts. For a meta path $M = E_{x_1} \xrightarrow{A_{x_1,x_2}} E_{x_2} \xrightarrow{A_{x_2,x_3}} \cdots \xrightarrow{A_{x_l,x_{l+1}}} E_{x_{l+1}}$, where $x_y \in [1, K]$, $y \in [1, l]$, following notations in Sun et al. (2011), we say a path instance $p_{e_1 \rightsquigarrow e_{l+1}} = (e_1 e_2 \ldots e_{l+1})$ between e_1 and e_{l+1} follows meta path M, if $\forall i$, $e_i \in E_{x_i}$ and each edge $\langle e_i e_{i+1} \rangle$ belongs to each relation $A_{x_i,x_{i+1}}$ in M.

Given a meta path M which encodes the correlation contexts, we design the following measure to compute $R(e_q, e, M)$, which denotes the relevance score between e_q and e based on M:

$$R(e_q, e, M) = \sum_{\forall e', p_{e_q \rightsquigarrow e'} \in M'} R(e_q, e', M') \times w_{e',e} \tag{4}$$

where $e_q \in E_{x_1}$, $e' \in E_{x_l}$, $e \in E_{x_{l+1}}$, $M' = E_{x_1} \xrightarrow{A_{x_1,x_2}} E_{x_2} \xrightarrow{A_{x_2,x_3}} \cdots \xrightarrow{A_{x_{l-1},x_l}} E_{x_l}$ and edge $\langle e' e \rangle \in A_{x_l,x_{l+1}}$. If meta path M is an empty path, i.e., $l = 0$, then we define $R(e_q, e, M) = 1$ if $e_q = e$, and $R(e_q, e, M) = 0$ otherwise.

When we want to use multiple meta paths $M_1, \ldots, M_k$ as correlation contexts, we can compute $R(e_q, e)$ as a linear combination of the relevance score over each meta path M_i:

$$R(e_q, e) = \sum_{i=1}^{k} \theta_i R(e_q, e, M_i) \tag{5}$$

where meta paths with higher weight θ_i are considered to encode more important correlation context. The meta path weight θ_i can be learned in a supervised manner (Lao and Cohen 2004), such as giving higher weights to certain meta paths so that entities that are labeled to be strongly relevant could have higher relevance scores. This problem is however out of the scope of this work. In this work, we manually tune the weights of meta paths and put our focus on the computation of $R(e_q, e, M)$.

It is not hard to derive that the relevance measure defined in Eq. 4 is equivalent to the following, which is efficient to compute:

$$R(e_q, e, M) = \sum_{p_{e_q \rightsquigarrow e} \in M} \left(\prod_{\langle e_i e_j \rangle \in p_{e_q \rightsquigarrow e}} w_{ij} \right) \tag{6}$$

where $\langle e_i e_j \rangle$ denotes any edge belonging to path instance $p_{e_q \rightsquigarrow e}$, and w_{ij} is the weight of the edge.

According to the above analysis, a pair of heterogeneous entities will have a high relevance score if: (1) they are strongly connected to other strongly relevant entities following selected meta paths; or (2) they are connected by paths with high weights following selected meta paths. This intuition clearly expresses the correlation context encoded in meta paths. For example, clindamycin hydrochloride is considered to be strongly relevant to acne because they are connected by many highly weighted paths following meta path "Drug–Compound–Disease" (such as "clindamycin hydrochloride–azelaic acid–acne"), and also because both clindamycin hydrochloride and acne are strongly relevant to many compounds in the middle (such as azelaic acid), which are will encoded in the edges and paths.

Our proposed measure has the good symmetric property, i.e., $R(e_q, e, M) = R(e, e_q, M^{-1})$, where M^{-1} denotes the reverse meta path of M (Sun et al. 2011). Moreover, it fully takes into account the carefully-designed weight of the edges (and in turn the paths) instead of just doing path counting (Sun et al. 2011; Shi et al. 2012), therefore well utilizing the rich contexts encoded in the paths. This design is also consistent with the weight of meta paths defined in Eq. 3.

7 Experiments

In this section, we empirically evaluate the effectiveness of our proposed framework for estimating the relevance between heterogeneous entities. We also present a prototype drug search engine built based on the framework proposed in this paper. We begin with the experimental setup.

7.1 Experimental setup

In order to evaluate the relevance estimation results generated by different algorithms, we sampled 199 unique drug–disease pairs from FDA's orange book[7] as the ground truth for the therapeutic relationships between drugs and diseases. We chose the therapeutic relationship as testing cases because it is one kind of strong relevance largely supported by the MEDLINE data. While sampling, we emphatically avoid those well-known drugs, as their relevance can be easily captured by their large number of co-occurrences with diseases. The co-occurrence distribution of our ground truth drug–disease pairs is illustrated by Fig. 4. It can be observed that most of the drugs that known to treat certain diseases co-occur rarely with the disease in the text (typically no more than 10 times out of the 20 million abstracts). Therefore, the relevance relationship that we want to discover is really hidden in the text and can hardly be discovered by simply counting the raw number of co-occurrences or by natural language processing techniques.

[7] http://www.accessdata.fda.gov/scripts/cder/ob/default.cfm. Among all the relevance relationships between different types of biological entities, we show the discovery results of the therapeutic relationships as an example since the results are easy to be evaluated by referring to FDA's orange book.

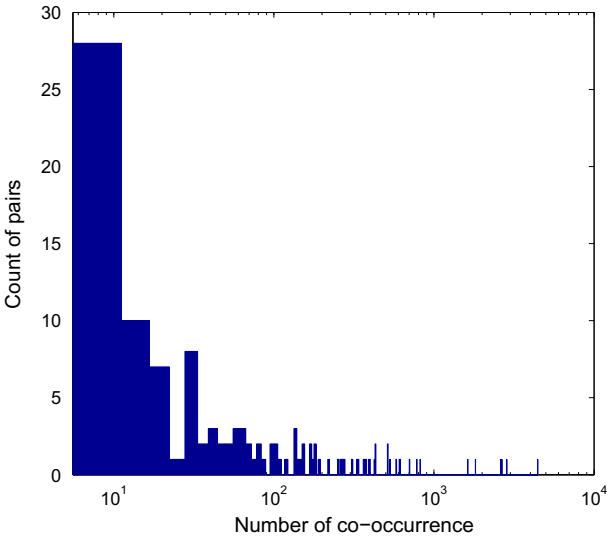
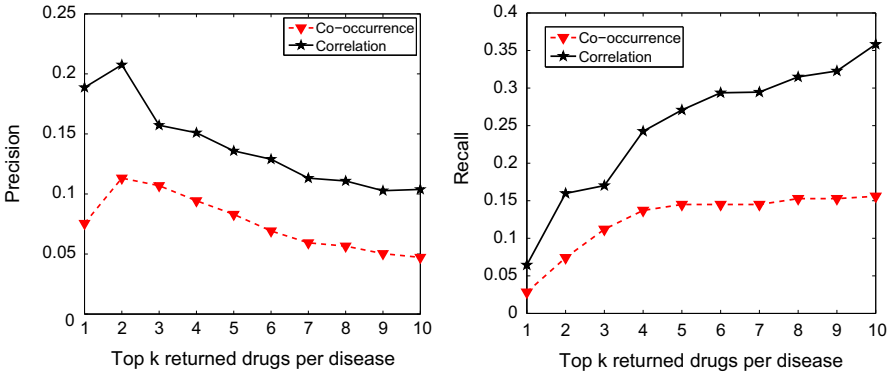


Fig. 4 Histogram of # of times that the ground truth drug–disease pairs co-occur in text corpus $\mathcal{D}$

Given a disease, all drugs in the database can be ranked according to the relevance scores, denoting how likely each drug is relevant to the disease. It is tricky to judge the "correct" returned drugs as we only have ground truths for the therapeutic relationship, not for strong relevance in general. To evaluate the correctness of a returned drug, not only will the drug be compared with ground truths (for Recall), but also the reasoning entities will be manually checked by human experts to see if the inferred relationship falls in the treatment category (for Precision). We use the standard precision, recall and Mean Average Precision (MAP) (Manning et al. 2008) to evaluate the results for our problem. Precision is defined as the # of drugs that can treat the queried disease based on human evaluation by the # of returned drugs. Recall is defined as the # of drugs in ground truth divided by the # of returned drugs. Given a disease, let r_i be the judgement score of the drug ranked at position i, where $r_i = 1$ if the drug is known to treat the disease and $r_i = 0$ otherwise. Then we can compute the Average Precision (AP):

$$AP = \frac{\sum_i r_i \times \text{Precision@}i}{\#\text{ of drugs known to treat the disease}}$$

And MAP is the average of AP over all the diseases in our labeled set. We do not use Normalized Discount Cumulative Gain (NDCG) to measure the performance, since we can only manually judge whether a drug can or cannot treat the given disease, but do not have knowledge about the levels, that is, about how much one drug can treat one disease.



(a) Precision **(b)** Recall

Fig. 5 Compare *correlation* to *co-occurrence*

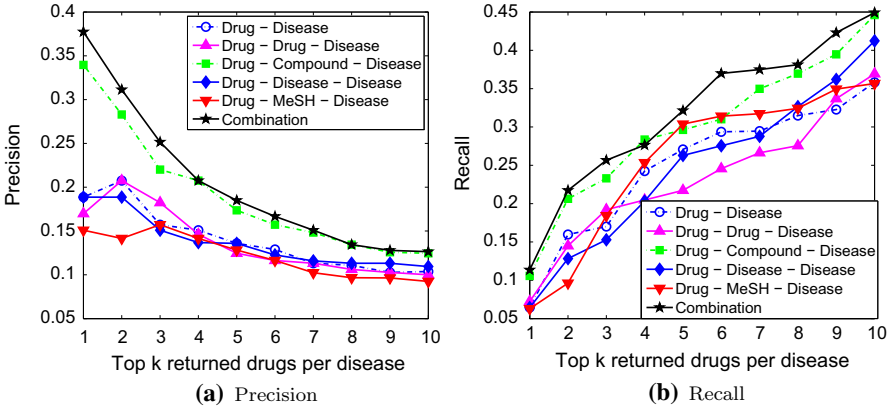
7.2 Correlation weight evaluation

We first evaluate the effectiveness of our proposed correlation weight function comparing to the raw co-occurrence count for entity correlation graph construction. As mentioned before, the correlation weight function could be used as one naive solution of $R(e_q, e)$, where e_q is the query disease and e is one drug. We show the average precision curves and recall curves of the co-occurrence based method (denoted by *co-occurrence*) and correlation weight function based method (denoted by *correlation*) with regarding to the top number of returned drugs per disease in Fig. 5. We show the precision and recall measures for the top 10 returned drugs per disease since the top 10 returned results are the most important to the user and largely affect the user experience. The MAP scores for *correlation* and *co-occurrence* are 0.216 and 0.118, respectively, measuring the performance over the entire ranking list. The correlation weight based method performs much better than the raw co-occurrence based method on all the above evaluation metrics, meaning that our design of the correlation weight function in Eq. 2 is reasonable to capture the direct correlation between two entities. So, we use Eq. 2 instead of the raw co-occurrence to construct the entity correlation graph.

7.3 Comparing different meta paths

We selected top 5 meta paths using Algorithm 1 (listed in Sect. 5.2). Based on a single meta path, we can find relevant drugs. We can also perform a weighted combination of the results generated by multiple meta paths. How to combine the results of different meta paths or how to set the weight for each path during combination is a difficult problem, which is left for future study. In this paper, we manually tuned the weights and picked up the weights with the best performance.

Based on EntityRel, we compare the retrieval performance of individual meta paths and their combination. We show the average precision curves and recall curves



(a) Precision **(b)** Recall

Fig. 6 Compare different meta paths and their combination in Precision/Recall based on EntityRel

Table 2 Compare different meta paths and their combination

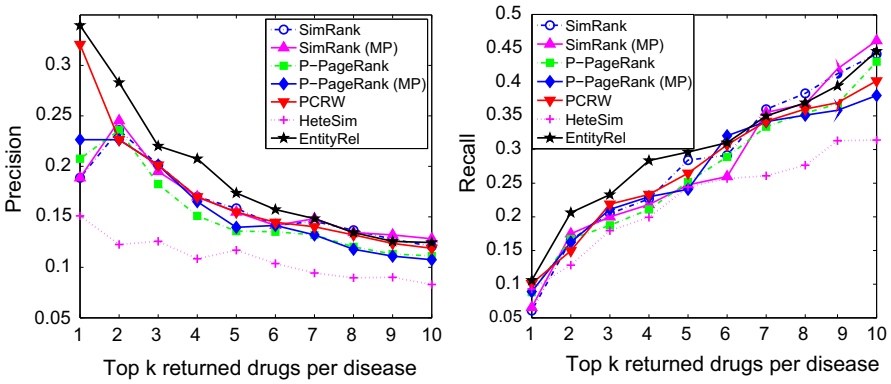
Meta path	MAP
Drug–disease	0.216
Drug–drug–disease	0.218
Drug–compound–disease	0.276
Drug–disease–disease	0.216
Drug–MeSH–disease	0.203
Combination	0.290

of various meta paths and their combination in Fig. 6, where the manually selected weights (similar to weights tuned by cross-validation) for the top 5 meta paths (shown in Sect. 5.2) are 0.01, 0.1, 0.78, 0.1, 0.01, respectively. Generally, we can see that combining the results generated by different meta paths performs equal or better than any single meta path, especially in the top few returned drugs. The MAP score comparison of different meta paths and their combination is shown in Table 2, where we can see the combination method achieves the highest MAP score, indicating the best overall performance. Among the results generated by one single meta path, path "Drug–Compound–Disease" performs the best.

Remember the MAP score of the *correlation* method in the previous subsection is 0.216, which is much lower than both the combination method and the best result generated by one single meta path. This indicates that employing the top meta paths can generate better results than simply using the *direct* correlations between drugs and diseases as the relevance estimation, since the former encodes the interrelationships between multi-typed entities in a structured way.

7.4 Comparing different methods on the same heterogeneous entity correlation graph

As mentioned before, the following state-of-the-arts can be used to estimate relevance between two homogeneous or heterogeneous entities. We adapted them on the same



(a) Precision **(b)** Recall

Fig. 7 Compare EntityRel to related work

heterogeneous entity correlation graph $\mathcal{G}'$ generated by top 5 meta paths for a fair comparison.

- Personalized PageRank (Jeh and Widom 2003). The damping factor is set as 0.9. By ignoring the type difference among entities and edges, it can be run on two different graphs in our scenario: (1) the original correlation graph $\mathcal{G}$, named P-PageRank; and (2) the graph $\mathcal{G}'$ which only contains top 5 meta paths, denoted by P-PageRank (MP).
- *SimRank* (Jeh and Widom 2002). Damping factor is set to 0.8. Has two versions similarly: SimRank on $\mathcal{G}$ and SimRank (MP) on $\mathcal{G}'$.
- *HeteSim* (Shi et al. 2012) run on its best meta path.
- Path Constrained Random Walk (*PCRW*) (Lao and Cohen 2010) run on its best meta path.

HeteSim, PCRW and our method EntityRel are all based on the best meta path selected from the top 5 meta paths. Although it is verified in the previous subsection that combining the results generated by multiple meta paths performs better than a single meta path, how to combine multiple meta paths without any label information is still an unsolved problem left for future study. Therefore, we simply run HeteSim, PCRW and our method on each of the top 5 meta paths and choose the best results for comparison.

It is worth noting that both original SimRank and original HeteSim work on binary graphs only, considering whether two nodes are connected or not and ignoring the weight on the edges. We tried the original versions on the binary correlation graphs without using the weighted edges, and found that they performed rather poorly. Therefore, we show these two methods' results on the weighted correlation graph only.

From the average precision curves and recall curves shown in Fig. 7, we can see that our EntityRel model leads the pack, especially when the number of returned drugs is small. PCRW performs the second best. Another observation is that SimRank performs similarly on the complete correlation graph $\mathcal{G}$ and the graph only containing selected meta paths $\mathcal{G}'$, and so does P-PageRank. This indicates that while reducing the time and space complexity largely, our top 5 selected paths capture most of the

Table 3 Compare EntityRel to related work in MAP

Algorithm	MAP
SimRank	0.251
SimRank (MP)	0.254
P-PageRank	0.245
P-PageRank (MP)	0.244
PCRW	0.253
HeteSim	0.204
EntityRel	0.276

useful information in the original graph. The MAP scores of different algorithms are shown in Table 3. We can see our EntityRel is still the best, achieving 8.66 % relative MAP score improvement over the second best algorithm. This indicates the reliable performance of EntityRel over the entire ranking list of returned drugs.

7.5 Discussion on the computation time

All the experiments in this section were conducted on a PC with 2.67 GHz CPU and 48 GB memory. And all the algorithms are implemented in MATLAB. The correlation graph construction step is computed offline, and is the same for all the relevance estimation algorithms. After graph construction, given one disease as the query, the path-based algorithms (EntityRel, HeteSim and PCRW) can compute the relevance scores for all the drugs in the database within 0.03 second, while P-PageRank and SimRank take minutes to hours to process one query. In this way, we can see our method EntityRel is an efficient approach for real applications such as online search engines for biomedical entities.

7.6 Prototype drug search engine

In addition to the theoretical contribution and empirical analysis presented above, we also implement a prototype drug search engine inside IBM. Note that rather than any structured data, the only data set used by this search engine is the biomedical paper corpus. Figure 8 shows a real example in our demo system, where a user submits a disease "acne"[8] and searches for strongly relevant drugs. All the top 10 returned results are FDA-approved drugs for treating acne. Specifically, the 10th drug "clindamycin hydrochloride" only co-occurs with "acne" and its synonyms 5 times in more than 20 million MEDLINE articles, which cannot be discovered by simple co-occurrence methods easily. Note that the correctness of strong relevance depends on the reasoning entities in the paths used to discover the relationship. All the five reasoning compounds (nadifloxacin, azelaic acid, doxycycline hyclate, minocycline, dapsone) in the paths that contribute most to this discovery result clearly indicate that the strong relevance found by us between "clindamycin hydrochloride" and "acne" is a valid therapeutic

[8] The hit disease "acne vulgaris" is its synonym.

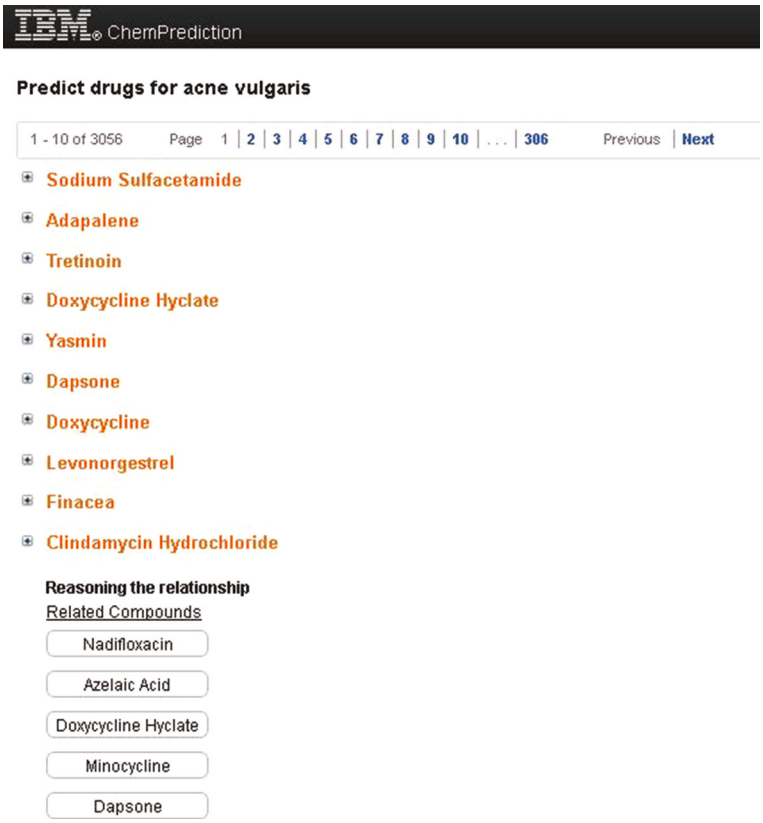


Fig. 8 Drug search engine demo

relationship. On the contrary, if we use similar contexts to reason the relationship of "vitamin A" (co-occur with acne 22 times) or "Insulin" (co-occur with acne 21 times) with "acne", the relationship will be wrong despite that these two drugs are relevant to disease "acne" in other ways. For example, to treat acne, large doses of vitamin A must be given, which then results in vitamin A toxicity; acne has an effect of insulin resistant. These relationships have to be detected by other correlation contexts, such as "symptom"-typed entities. In this way, when we judge the correctness of the discovered strong relevance, we utilize the set of reasoning entities involved in the relevance discovery. This drug search engine has been presented to medical experts, who have evaluated the search results and agreed that the drugs returned by the search engine are indeed strongly relevant to the query disease.

8 Conclusions

In this paper, we propose a framework to solve the critical problem of discovering strong relevance between heterogeneous typed entities from massive biomedical text data, which is challenging due to the unstructured and noisy nature of text. To achieve

the goal, we build an entity correlation graph from the text data, which offers rich semantic contexts for entity relationships. In our approach EntityRel, two entities of heterogeneous types are strongly relevant if they are linked closely to each other or to other strongly relevant entities along paths following the selected patterns. EntityRel is effectively evaluated on the examples of Drug–Disease relevance discovery over 20M+ MEDLINE abstracts. Although we focus on biomedical data in this paper, our approach EntityRel is generic enough to be applied in other domains. For example, given a social network containing friendship relationships between users, and check-in relationships between users and restaurants, we can apply EntityRel to discover strong relevance between users and restaurants, which could be used to predict restaurants that each user likes. For future work, one interesting problem is to learn the weights of different meta paths. Another interesting direction is to study the theoretical relations between EntityRel and other random walk models.

Acknowledgments Research was sponsored in part by the Army Research Lab, under Cooperative Agreement No. W911NF-09-2-0053 (NSCTA), National Science Foundation IIS-1017362, IIS-1320617, IIS-1354329, HDTRA1-10-1-0120, and NIH Big Data to Knowledge (BD2K) (U54).

References

Aleman-Meza B, Halaschek-Wiener C, Arpinar IB, Sheth AP (2003) Context-aware semantic association ranking. In: Semantic Web and Databases, pp. 33–50

Anyanwu K, Maduko A, Sheth AP (2005) Semrank: ranking complex relationship search results on the semantic web. In: WWW, pp. 117–127

Anyanwu K, Sheth AP (2003) P-queries: enabling querying for semantic associations on the semantic web. In: WWW, pp. 690–699

Coulet A, Garten Y, Dumontier M, Altman R, Musen M, Shah N (2011) Integration and publication of heterogeneous text-mined relationships on the semantic web. J Biomed Semant 2(Suppl 2):S10

Eppstein D (1998) Finding the k shortest paths. SIAM J Comput 28(2):652–673

Guan Z, Wang C, Bu J, Chen C, Yang K, Cai D, He X (2010) Document recommendation in social tagging services. In: WWW, pp. 391–400

Gunther E, Stone D, Gerwien R, Bento P, Heyes M (2003) Prediction of clinical drug efficacy by classification of drug-induced genomic expression profiles in vitro. Proc Natl Acad Sci 100(16):9608

Jeh G, Widom J (2002) Simrank: a measure of structural-context similarity. In: KDD, pp. 538–543

Jeh G, Widom J (2003) Scaling personalized web search. In: WWW, pp. 271–279

Lao N, Cohen WW (2004) Relational retrieval using a combination of path-constrained random walks. Mach Learn 81:53–67

Lao N, Cohen WW (2010) Fast query execution for retrieval models based on path-constrained random walks. In: KDD, pp. 881–888

Manning CD, Raghavan P, Schütze H (2008) Introduction to information retrieval. Cambridge University Press, Cambridge

Ramakrishnan C, Mendes P, Wang S, Sheth A (2008) Unsupervised discovery of compound entities for relationship extraction. Knowledge Engineering: Practice and Patterns pp. 146–155

Searls D (2005) Data integration: challenges for drug discovery. Nat Rev Drug Discov 4(1):45–58

Sen S, Vig J, Riedl J (2009) Tagommenders: connecting users to items through tags. In: WWW, pp. 671–680

Sheth AP, Aleman-Meza B, Arpinar IB, Bertram C, Warke YS, Ramakrishnan C, Halaschek C, Anyanwu K, Avant D, Arpinar FS, Kochut K (2005) Semantic association identification and knowledge discovery for national security applications. J Database Manage 16(1):33–53

Shi C, Kong X, Yu PS, Xie S, Wu B (2012) Relevance search in heterogeneous networks. In: EDBT, pp. 180–191

Sun Y, Han J, Yan X, Yu PS, Wu T (2011) Pathsim: meta path-based top-k similarity search in heterogeneous information networks. PVLDB 4(11):992–1003

Yan S, Spangler WS, Chen Y (2011) Cross media entity extraction and linkage for chemical documents. In: AAAI

Yin D, Xue Z, Hong L, Davison B (2010) A probabilistic model for personalized tag prediction. In: KDD, pp. 959–968