Materials Science Literature-Patent Relevance Search: A Heterogeneous Network Analysis Approach

Pingjie Tang*, Jed Pitera†, Dmitry Zubarev†, Nitesh V. Chawla*
*Dept. of Computer Science and Engineering
University of Notre Dame
Notre Dame, USA
{ptang, nchawla}@nd.edu
†IBM Research - Almaden
San Jose, USA
{pitera, dmitry.zubarev}@us.ibm.com

***Abstract*—In recent decades, materials science literature and patents have grown exponentially. This has also contributed to an ever-growing challenge whether the literature is current, as there can be a gap between when the patent was filed and when it was approved. Moreover, it is difficult to ensure that a patent cites the appropriate prior art due to variety and volume of materials science data, especially when it is in two separate sources that have different curation mechanisms and purpose — publications and patents. The existing relational database schema, generally used to store publications, also presents challenges given the strict tabular schema, which may not be appropriate for organizing and querying highly interconnected information about materials in these publications and patents. For example, elements are chemically combined to form a compound, which can then be converted to other compounds via chemical reactions. Furthermore, relational database is not designed for handling combining data from multiple sources and with various formats, thus it makes discover relevance between publications and patents become difficult.**

In order to explore an alternative approach to represent materials data and combine data from multiple sources into the same repository, in this work, we propose a solution to integrate data from Open Quantum Materials Database (OQMD) and patent data from USPTO[1] database into a network and named it heterogeneous materials information network (HMIN). We generalize prior work which based on using meta path-based topological features to explore the network, and we propose features to identify network noise and investigate relatedness between different-typed objects to meet our application needs.

We built several machine learning models by using these features to explore relevance between materials science publications and patents. Experiment results show that HMIN can help researchers effectively discover related publications and patents originally kept in different sources.

Our work exhibits to materials community a new way of appropriately representing materials data and discovering connections between data from multiple sources.

Keywords*—*heterogeneous information network; meta path; relevance search; data integration

I. Introduction

The evolution of materials science research has accompanied massive amounts of both structured and unstructured data. Big data driven materials science has drawn more attentions over last few years. Several review articles [1], [2], [3] identify common and distinctive features in materials data, such as data generated by experiments, simulations, and also literature. Materials data tends to be very heterogeneous and is distributed across various sources. Lacking standard and widely adopted protocols to manage data and evaluation metrics to estimate data quality has left uncertainties in materials data.

In this paper, we are especially interested in two unique characteristics of the materials data. We first observed that the materials data has a graph-like nature due to data is interconnected with each other. For example, chemical compounds and elements interact with each other through chemical reactions, or hidden relations exist among chemical compounds which share common physical properties, practical applications, and environmental or economic attributes. These hidden or latent relationships might help suggest new uses for a material, or new connections between different types of materials. Secondly, materials data exists in various formats, and similar data are often encoded in many ways. It is difficult to combine data from multiple sources, understand and reuse existing data, and find associated metadata [4]. Those distinctive characteristics impose design of solutions to appropriately represent and store materials data, thereby enabling managing connected data, working with complex queries, and making access, update and analyze data easily and quickly.

Most materials data is stored in relational databases, such as OQMD[2], ICSD[3] and MatNavi[4]. Relational storage system has deficiencies in coping with data variations. It requires predefined database schemas before loading data, and the whole database thus becomes cumbersome when slight modifications made to the schema. Furthermore, it accommodates data inside tables, with rows denoting data entities and columns representing data attributes. For poorly designed database, information

[1] http://patft.uspto.gov/netahtml/PTO/search-bool.html

[2] http://oqmd.org/

[3] https://icsd.fiz-karlsruhe.de/search/index.xhtml;jsessionid=009CA4105F40077B535C24

[4] http://mits.nims.go.jp/index_en.html

978-1-5090-5004-8/17 $31.00 © 2017 IEEE
DOI 10.1109/DSAA.2017.8

about individual object is scattered in multiple tables due to normalization considerations [5], and people have to rely on heavy-join operations to extract complete data information and mutual interactions to fulfill complex queries. In addition, SQL (Structured Query Language) syntax becomes complex and large as join operations increase. Inabilities of the relational database to handle data variations and interconnections posit a need for alternative approaches that can efficiently store and combine both relational and non-relational data about materials.

A practical problem that materials science researchers usually encounter is to discover related patents and publications. Due to the variety and volume of materials science data, it is difficult to ensure that a patent cites the appropriate prior art. Algorithms to connect a patent with relevant literature citations based on mention of a common material would be useful. In most cases, however, patent and publication data are stored in different data repositories. Our paper aims to develop a proper data model to combine data from multiple sources and of different formats. Meanwhile, we investigate approaches to explore correlations between publications and patents to meet application requirements.

As a schema-free data model, network represents interconnected data more naturally without requiring researchers to force fit their complex structured data into tabular data formats ill-suited to their needs. Instead of storing data into relational tables, data objects are represented as nodes and data relations are mapped to links between nodes in the network. Network mechanism benefits from the native index-free adjacency property, which turns complex join operations into rapid graph traversals [6], [7]. When we perform a graph query over the network, graph traversal can efficiently explore the network without looking up indices and is not completely dependent on the network size. Assume that we are looking for which papers study silicon oxide, graph search first locates silicon oxide node by using certain graph search method (depends on network design), the total cost of searching relevant papers is only proportional to the size of paper nodes linking to the silicon oxide node. Searching cost stays the same if the number of related papers remains unchanged even if the entire network size increases. Without predefined schema, network can efficiently and flexibly integrate data and relations from multiple sources by easily creating nodes, edges and attributes. Most real world data is heterogeneous and usually has multiple data types and relation types. In relational database, data records within different tables belong to distinct types, and foreign key constraints describe various types of relations among tables. Recent years have witnessed a proliferation of large scale of heterogeneous network applications across multiple domains [8], [9], [10], [11], [12], [13], [14]. However, there are few works integrating data from multiple databases into a heterogeneous information network. In this paper, we proposed a method for building a heterogeneous information network (HIN) by integrating partial data from a materials relational database OQMD with data from US patent database. Different from homogeneous networks, HIN usually contains rich and complex semantics, therefore different paths between a pair of nodes may contain distinct semantics. Meta path is a meta-level description of a path [11], [15]. It denotes concatenated relations between two data objects, we leveraged it to capture rich semantics of HIN. For example, paths between two compositions, which are represented as "composition$_x$-structure$_k$-composition$_y$" and "composition$_x$-structure$_a$-publication$_m$-structure$_b$-composition$_y$", respectively follows meta paths "composition - structure - composition" and "composition - structure - publication - structure - composition". The first meta path describes a pair of compositions share the same chemical structure, while the second one interprets two compositions possessing different structures, and both structures are studied in a publication. Meta path enables materials researchers to search for materials data over HIN using meta paths that are most close to their requirements. It is not difficult to find meta path help people discover appropriate query answers and narrow the search space according to query semantic meanings.

To explore relationships between materials science patents and publications, we trained several supervised learning models using various features to predict objects relevance. We designed features to identify network noise induced by keyword-based approach, which extracts relations between patents and OQMD chemical compositions. Distinct from most HIN applications [11], [15], [9], our work aims to evaluate similarities between different-typed objects. We designed features combining Hetesim [16], which evaluates the relatedness between objects with different types with heterogeneous network topological structure information to quantify publication-patent relevance. The performances of proposed methods were evaluated using precision, recall and F1-score. Our results demonstrated models trained with different feature combinations calculated based on the heterogeneous network entities and topology can effectively identify network noise and predict objects relatedness.

To our best knowledge, at the time of writing, our work is the first effort of building a materials science heterogeneous information network and providing solutions to predict network objects relevance. The main contributions of this paper are as follows:

1) Provide a new and effective paradigm of data organization and analysis approach which is different from traditional solutions developed in the materials community.
2) Integrate data from relational materials database and patent data from patent database into a heterogeneous materials information network.
3) Explore the network via different feature combinations designed based on network topology and data, and employ machine learning models to discover relevant objects.

The rest of this paper is organized as follows. In section II we introduce the problem definition. Section III describes methods we used to construct the network and discover the objects relevance. Experiments setting and results will be discussed in section IV. Section V lists related works and section VI contains the conclusion.

II. Problem Definition

In this section, we define the heterogeneous materials information network construction and the relevance search task. Meanwhile, we introduce HIN meta path and topological features as preliminary knowledge.

Our goal is to find correlations between objects from disparate data sources, specifically, to discover relevance between publications in the OQMD and patents from USPTO database. Based on the patent content, additional publication references can be indirectly suggested through the materials mentioned in the patent. These additional references can enrich and provide content for the patent citations.

A. Heterogeneous Materials Information Network (HMIN)

We integrated partial data from OQMD v1.0 database with relevant patent data from USPTO by building a heterogeneous information network. More detailed information about OQMD database is discussed in [17]. OQMD v1.0 has 71 relation tables but we only extracted nine tables out of them in order to fit our application. Each table stands for one node type in the network. Hence HMIN originally included node types: "authors", "compositions", "compositions_element_set", "elements", "journals", "publications", "publications_author_set", "spacegroups" and "structures". Considering "compositions_element_set" and "publications_author_set" are two junction tables that resolve many-to-many relationships. We converted data from these two tables into network edges between corresponding object nodes. Foreign key relationships were mapped into links between data entities accordingly (See Sec. III).

In order to expand the current network built based on OQMD with patents, we invited materials scientists to identify the connections between the current network and patents. A set of patents (identified by patent numbers) were extracted by considering relationship between patents and chemical composition objects in the network. In addition, International Patent Classification (IPC) and citations associated with each patent were included into the network. Four patent related node types: patent, major IPC, minor IPC and citations were added to the current network. Figure 1 shows HMIN network schema. Nodes created for relation table attributes are not shown in this figure due to limited space.

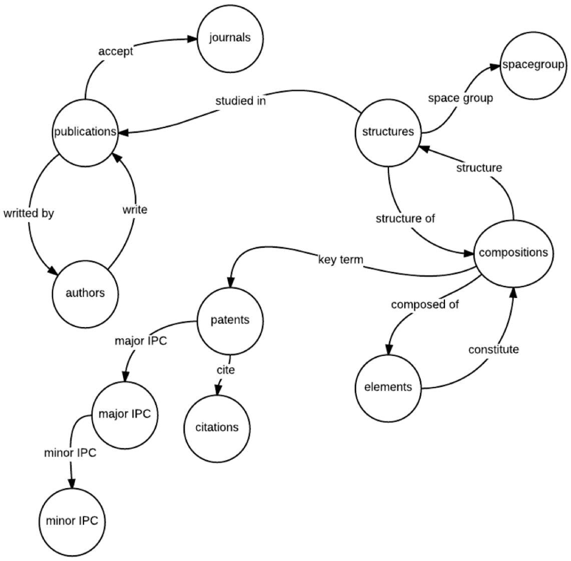


Fig. 1: HMIN Network Schema

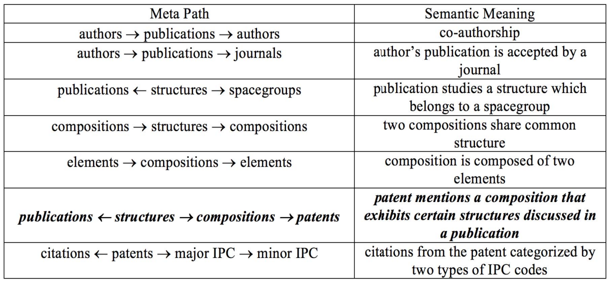
B. Publication-Patent Relevance Search

Our second task is to perform a relevance search over HMIN to find related publications and patents. First, we need to find entire set of paths connecting publication and patent nodes that follow certain type of meta paths, since the relevance search needs to be guided by query semantics to avoid providing unrelated results. Querying "chemical structures sharing the same space groups" is different from querying "chemical structures that are studied in publications accepted by some journal". We then investigate meta path-based topological features and other features designed based on our application needs to build supervised learning models to explore whether materials literature and patents are truly related.

1) Meta Path In HIN: A heterogeneous information network is usually defined as a directed graph G = (V, E), each node v ∈ V belongs to a particular node type of the node types set K, likewise, each edge e ∈ E can be mapped to a certain edge type of the edge types set R. The ***network schema*** is a network meta template of G, which is a directed graph defined over node types K and edge types R, denoted as G_S = (K, R). The network schema of HMIN is shown in Figure 1.

DEFINITION 1. ***Meta Path***
A meta path ***P*** is a path defined over the network schema G_S, which defines a composite relation of adjacent links between starting object type K_1 and ending object type K_{t+1}. It is denoted in the form of $K_1 \xrightarrow{R_1} K_2 \xrightarrow{R_2} ... \xrightarrow{R_t} K_{t+1}$. Meta path is a "meta-level" description of a network because it consists of a sequence of node and edge types. We take a path connecting two chemical elements from OQMD for an example, "Ac1 Ag1 Ca2 $\xrightarrow{\text{has structure}}$ structure_id:1120157 $\xrightarrow{\text{space group}}$spacegroup_id:255 $\xrightarrow{\text{space group}^{-1}}$ structure_id:1383008 $\xrightarrow{\text{has structure}^{-1}}$ Ac1 Ag1 Ce2". Its corresponding meta path is "composition $\xrightarrow{\text{has structure}}$ structure $\xrightarrow{\text{space group}}$ space group $\xrightarrow{\text{space group}^{-1}}$ structure $\xrightarrow{\text{has structure}^{-1}}$ composition". This meta path matches entire set of actual paths with semantic meaning: Two chemical compositions possess structures belonging to the same space group. We call these paths following the meta path ***P*** as ***path instances*** of ***P***. R^{-1} is an inverse relation of R. The length of ***P*** is number of links on ***P***. Figure 2 lists a portion of meta paths under length 4 discovered in HMIN. It is noteworthy that we only chose meta paths of length 4 because those ones with length greater than 4 are considered to be too long to make a significant contribution [13]. The bold-faced italic row shows the meta path and its semantic meaning studied in this work. Different meta paths in the network can be enumerated using standard graph traverse algorithms such as breadth first search (BFS) or depth first search (DFS). Meta path is defined on the network schema, which is a much smaller graph than the actual network, thus exploring meta path is very efficient.

2) Network Topological Features: Any pair of objects sitting on both ends of a meta path are connected in a semantic way, however, it is not confident to say they are correlated highly with each other. For example, two people posted pictures taken at golden gate bridge online. It can be



Meta Path	Semantic Meaning
authors → publications → authors	co-authorship
authors → publications → journals	author's publication is accepted by a journal
publications ← structures → spacegroups	publication studies a structure which belongs to a spacegroup
compositions → structures → compositions	two compositions share common structure
elements → compositions → elements	composition is composed of two elements
publications ← structures → compositions → patents	***patent mentions a composition that exhibits certain structures discussed in a publication***
citations ← patents → major IPC → minor IPC	citations from the patent categorized by two types of IPC codes

Fig. 2: Meta paths under length 4 in HMIN

described by meta path: person $\xrightarrow{\text{post}}$ picture $\xrightarrow{\text{where}}$location $\xleftarrow{\text{where}}$ picture $\xleftarrow{\text{post}}$ person. It describes two people posted pictures taken at the same place, but it doesn't reflect relation between two people, they might know each other or they might not.

Heterogeneous topological features are often used to measure network objects correlation and can generate more accurate prediction results for various applications in HIN compared to homogeneous topological features [11], [15], [9]. Some frequently used topological features defined in homogeneous networks such as, common neighbors, Jaccard's coefficient and Adamic/Adar [18] are prone to observe paths or neighbor set patterns to determine object similarities. There are also works employ graph cliques as the network features being used in the graph classification task [19], [20] . However, such network features would not be applicable in the heterogeneous scenario because an object's neighbors can include multiple types of objects, and the paths between two objects could follow different meta paths and thus imply different semantics. In our work, we built multiple supervised learning models utilizing hybrid meta path-based topological features, which are path count, normalized path count, random walk and symmetric random walk to evaluate the relevance between publications and patents following the meta path *"publication - structure - composition - patent"*.

III. Methods

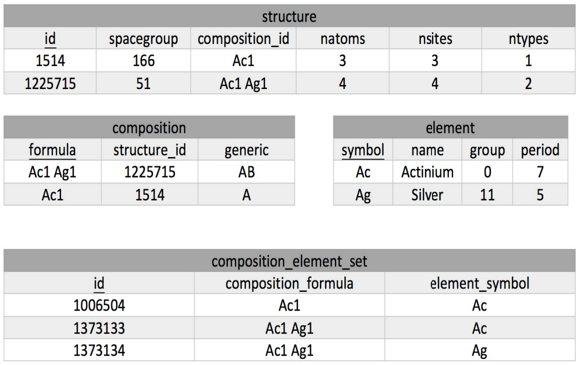
A. Heterogeneous Materials Information Network Construction

In this section, we discuss how to construct a heterogeneous network by combining data from distinct sources in detail. We first illustrate how to transform data from a relational database to a network. A relational database schema R is composed of (I) a set of relations (tables) r_1, ..., r_n. $r.A$ denotes attribute A of relation r, and X is an entry over r. Cell C is the unit where a row (entry) and a column (attribute) intersects. (II) Foreign key relations. We denote a foreign key relationship existing between relation r_x and r_y with FK_{xy}. (III) Many-to-many relations. Many-to-many relationships existing between data entities are resolved in junction tables and we denote it with *m2m* relation.

Given a relational database the heterogeneous network built based on it is a directed graph RG = (V, E). For entry X in the relation table r, we first mapped primary key field to a ***center node***, it uniquely identifies the entry X over r. Center node was labeled with the node attribute *"relation name: primary key: primary key value"*. Cell corresponding to each of remaining attributes in X was defined as ***attribute node*** pointing out of the center node. It reflects entry X identified by center node has an attribute field represented by an attribute node. Attribute node was labeled with *"relation name: attribute: attribute value"*.

An edge between nodes $v_i \in V$ and $v_j \in V$ is denoted as $e_{ij} = (v_i, v_j) \in E$, if one of the following holds: (I) v_i represents a center node and v_j is an attribute node, and both of them come from the same data entry X. (II) Foreign key relationship FK_{ij} exists between tables r_i and r_j. e_{ij} is the edge from center node v_i representing the entry containing the foreign key field in r_i to center node v_j representing the entry containing the primary key in r_j. (III) Many-to-many relationship exists between a primary key field in r_i and primary key field in r_j. It is mapped to a pair of edges with opposite directions between center node v_i and center node v_j.

Figure 3 shows an example of relational database which includes 4 tables. Primary key in each table were underlined. There are two foreign key relations: "structure.composition_id $\xrightarrow{\text{FK}}$ composition.formula" and "composition.structure_id $\xrightarrow{\text{FK}}$ structure.id". Junction table composition_element_set reflects presence of many- to-many relationships between composition and element objects. Figure 4 shows a network generated based on the relational database (see figure 3). (I) Node with bold circle indicates the center node of relation *r*. For example, "symbol" is the primary key of table element, and "Ac" is one of primary key values, the first row of element is mapped to a center node labeled with "element:symbol:Ac". (II) Attribute nodes are those ones pointing out of the center node. Attribute nodes surrounding the center node "structure:id:1514" indicate the entry containing primary key "1514" in table structure has four attributes: "spacegroup", "natoms"," nsites" and "ntypes" with respective attribute value associated. (III) Network edges are categorized into three types. Attribute edge represented with the solid line connects a center node and its attribute nodes. It represents entity-attribute relationship. Foreign key edge is denoted with dashed lines to describe foreign key relationship between tables in the database. Two foreign key relationships between relation composition and structure are converted to foreign key edges with edge type "FK:foreign key" in the network.



structure					
id	spacegroup	composition_id	natoms	nsites	ntypes
1514	166	Ac1	3	3	1
1225715	51	Ac1 Ag1	4	4	2

composition		
formula	structure_id	generic
Ac1 Ag1	1225715	AB
Ac1	1514	A

element			
symbol	name	group	period
Ac	Actinium	0	7
Ag	Silver	11	5

composition_element_set		
id	composition_formula	element_symbol
1006504	Ac1	Ac
1373133	Ac1 Ag1	Ac
1373134	Ac1 Ag1	Ag

Fig. 3: An example of relational database

The third edge type reflects the many-to-many relation-

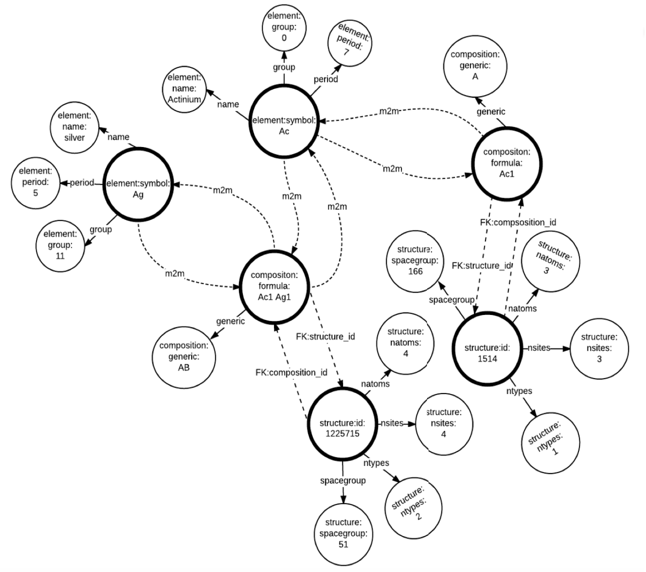


Fig. 4: An example of a HIN generated based on data from OQMD

ship. It is represented with dotted lines labeled with “m2m”. “composition_element_set” table in figure 3 accommodates many-to-many connections between tables “composition” and “element”, such information was consequently mapped to a pair of m2m edges linking composition and element-typed center nodes with inverse directions.

Work [21] focuses on converting a relational database to a graph database by employing a property graph model. It represents a table entry with a node and encapsulates key and non-key attributes with the node to form the node properties. Node also contains other table's attributes if a foreign key relation exists in between. Different from network construction strategy of work [21], our network structure needs to facilitate in exploring paths between objects in order to investigate objects relevance, relations hidden in the database have to be shown as a type of edges in the network. We considered foreign keys uniquely identify the relationship between tables, thus foreign key would not be shown as a part of node property, instead we created a foreign key edge. The same case applied to many-to-many relations. In addition, we were inspired by “subject-predict-object” triple used as the basic unit structure in RDF triple store to build the entry-attribute relation. Subject refers to the center node, object is the attribute node and predict refers to the edge type between center and attribute node. Such representation is very intuitive and can be interpreted semantically.

In order to integrate patents into current network, materials scientists used chemical compositions in the existing network as keywords to retrieve patents from USPTO database. Figure 1 shows newly added patent-related node types to the network. Composition-patent link is the bridge to combine data from distinct sources. Similarly, patent node plays the role of center node, IPC nodes and citations nodes are attribute nodes attached to it. Edges in between represent entity-attribute relationships aforementioned. The edge connecting a composition and the patent mentioning it are labeled with “key term”.

B. Impact Factors for Network Construction

There are some impact factors from either OQMD database end or patent extraction approach that make network construction become more challenging.

1) Duplicate Records: OQMD database contains a number of duplicate records across multiple tables. For example, the publications table has 210,999 tuples (entries), and structures table contains 1,542,049 tuples, however 133,510 (63%) of publications records and 361 (0.02%) structures records are repetitive. These redundant records should not be included in the network. Other types of duplicates in OQMD are also observed. For instance, some publication records actually refer to the same publication entity but written in different languages are kept as different records. Some publications of the same year published in the the same journal with identical volume number only have nuances in the titles, such as capitalized initial letters difference or with or without white spaces between words. In these scenarios, we only created one single node to keep one record written in English to maximally avoid network data redundancy.

2) Null Values in Foreign Key Field: Notwithstanding that it is a legitimate design decision to allow NULL-able foreign keys in the database, it results in problems while creating edges representing foreign key relations. For example, in a foreign key relation of OQMD, “compositions.structure_id $\xrightarrow{\text{FK}}$ structures.id”, each “compositions.structure_id” attribute field is filled in a NULL value. It made adding a link between node composition and node structure become infeasible. To tackle this problem, we manually added a reverse edge for each existing foreign key edge representing “structures.composition_id $\xrightarrow{\text{FK}}$ compositions.structure_id” relationship. We addressed the issue occurred in “structures.reference_id” attribute in the same way, it is a foreign key contains only NULL values and refers to the primary key “publications.id”. Lastly, foreign key correlations “structures.reference_id $\xrightarrow{\text{FK}}$ publications.id’ were completed by referring an intermediate table “entries” .

3) Patents Extraction: The approach we employed to extract related patents introduced noise to the network. We leveraged the keyword matching method to find connections between patent and chemical composition objects. Entire set of composition formula strings were extracted from OQMD as key terms to search matching patents. Patents that contain composition strings in the text were identified as relevant ones to the chemical composition. Accordingly, a patent node were created, and composition-patent edge were added to the network consequently. This information extraction method was very heuristic and inevitably introduced noise to the network. One primary source of noise stemmed from the presence of chemical formulas as either standard English language words or abbreviations in different domains. For example, the composition formulae for pure indium and pure boron are “I” and “B”, respectively. Both of these are common strings in patent text that may not necessarily refer to the materials in question.

C. The Publication-Patent Relevance Search Model

We formulated discovering related publications and patents over the network as a relevance search problem. In this section, we discuss how to construct three sets of features based on network topology and objects, and build several machine learning models using these features to perform the relevance search task.

Feature I is a hybrid of four heterogeneous meta path-based topological features extracted based on meta paths between a pair of objects. Feature I includes features of path count, normalized path count, random walk and symmetric random walk. They are extensions of homogeneous network topological features and can generate more accurate prediction results as compared to homogeneous topological features [15].

We have briefly discussed the network noise introduced by the heuristic way to extract patents. We came to realize that retrieving related patents to composition objects by merely checking whether patent text contains composition formula strings might lead to erroneous composition-patent connections. Some compositions only consist of single chemical element, such as "P", "Th", "Si", which might appear as a sub-string of a composition string such as: "AlP", "$Th(OH)_4$" and 'FeSi2'. A connection between a patent about the chemical compound (Na_2CO_3) and a composition composed of a single element (Na) can be wrongly added into the network. For each publication-patent path, we set a binary feature and named it ***pure material feature*** to identify if path includes an edge between a single-element composition and a patent. If it does, we set this feature to 1, otherwise, we set it to 0. **Feature II** was thus designed by including both Feature I and pure material feature. Unlike existing works that only leverage network topological features, an additional strength of Feature II is it also includes the network objects information (pure material feature) as well.

Existing studies focus on measuring similarity between objects of the same type [11], [15], such as author-to-author, publication-to-publication. Our work, however, aims to explore publication-patent connection. Therefore we designed **Feature III** by combining Hetesim [16] with Feature I in order to effectively evaluate the relatedness of heterogeneous objects.

Random Forest, Gaussian Naive Bayes and SVM with RBF (Radial Basis Function) kernel were trained using feature set I, II and III in determining relationships between publications and patents.

IV. Experiments and Results Discussion

In this section, we introduce basic information of the HMIN network. We report and discuss experiment results produced from classification models trained using different feature sets. Results evaluated by performance metrics show learning models utilizing heterogeneous network-based features can achieve satisfying predictive performance in measuring different-typed objects relatedness

A. Network

HMIN contains 11,614,523 nodes and 43,432,668 edges. Table I presents network nodes statistics. We only counted number of center nodes for each node type, given center nodes play the significant roles in network exploration. We used SNAP, a general-purpose network analysis library developed by Leskovec et al. [22] to build the network. SNAP is developed for efficiently manipulating massive and dynamic networks. Time complexity of key graph operations and memory usage of SNAP are lower than alternative approaches. Flexible network operations make the network properly and efficiently handle the data schema change issue that is usually can not be elegantly addressed in relational database. Given there are 77,489 publication nodes and 3,376 patent nodes in the network, it is not feasible to manually label data through randomly comparing each pair of patent and publication. Focusing on comparing publication and patent on paths following certain meta path will significantly reduce the search space and hence save manpower during data labelling process. Materials science researchers only need to investigate publication-patent path instances following the target meta path "publications - structures - compositions - patents". There were only 576 publication and 2,154 patent objects met such criteria. Namely, the only needed to compare 0.47% of the entire publication-patent pairs. 231 publication and patent pairs were eventually evaluated and labeled with score of 0, 1 or 2 to indicate definitely irrelevant, not determined or definitely relevant respectively. 120 pairs were identified definitely relevant, 74 pairs were identified irrelevant, and the rest 37 pairs were uncertain. Despite pairs of score one were still promising to be regarded as relevant pairs, we merged them with those of score zero as negative samples, and only treated pairs of score two as positive samples.

Node Types	Node Size
publications	77,489
journals	1,478
authors	53,808
structures	1,541,688
compositions	280,312
elements	112
spacegroups	230
patents	3,376
major IPC	276
minor IPC	1,829
citations	36,284

TABLE I: HMIN Network Nodes Statistics

B. Results and Discussion

In this section, we compare performances of learning models by using different features discussed in section III C, and report the average performance in terms of precision, recall, F1 score.

We first built supervised learning models by using Feature II to identify presence of network noise. We used results generated by using Feature I as the baseline and experiment results were summarized in table II. Three classifiers yielded better performances by incorporating pure material feature in topological features. In terms of precision, Gaussian Naive Bayesian (NB) classifier increased precision by 15%, Random Forest (RF) increased it by 7.86%, and SVM increased it by 10.15%. Recall values were improved roughly 20% via NB, 43.95% improvements were contributed by RF, and SVM boosted the value by 38.35%. F1 score is a harmonic mean of precision and recall. NB improved it by 19.58%, RF increased it by 26.64% and SVM enhanced F1 by 26.87%. Promising prediction results showed in table II demonstrates Feature II can be regarded as a discriminative feature in HMIN to identify erroneous network links and directly affect relevance search results. Figure 5 shows a more intuitive results comparison between the baseline and Feature II.

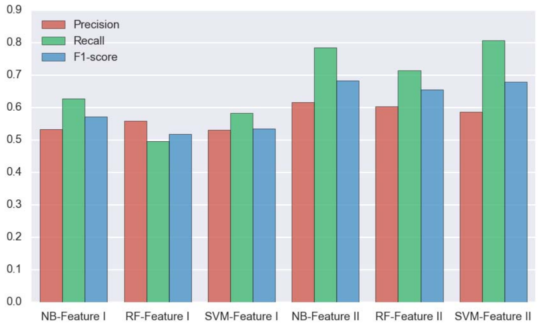


Fig. 5: Performance Comparison Between Baseline (Feature I on HMIN) And Feature II on HMIN

Feature set	Method	Precision	Recall	F1
Feature I	NB	0.534(0.042)	0.628(0.081)	0.572(0.044)
	RF	0.560(0.069)	0.496(0.131)	0.518(0.104)
	SVM	0.532(0.028)	0.584(0.227)	0.536(0.129)
Feature II	NB	0.616(0.080)	0.784(0.072)	0.684(0.072)
	RF	0.604(0.096)	0.714(0.081)	0.656(0.087)
	SVM	0.586(0.061)	0.808(0.074)	0.680(0.066)

TABLE II: Predictive Performance on HMIN By Introducing Pure Material Feature

Results obtained by involving Feature II showed the impacts brought by the network noise on learning models performance. In order to further improve network quality thereby gaining better performance for relevance prediction, we eliminated erroneous links between single-element composition and patent from HMIN, and obtained the trimmed network as **HMIN II**. Because some publication-patent paths for training will be cut off due to removing erroneous edges, afterwards, we had 51 negative samples and 115 positive samples left in the labeled dataset. We applied SMOTE [23] to alleviate the slight data imbalance issue. In order to show the performance variation incurred by network quality change, we built learning models by leveraging Feature I and respectively applied them to HMIN and HMIN II. Table III describes comparison results generated in different networks. After improving the network quality, NB achieved 38.2% improvement on precision, RF and SVM improved precision by 30.71% and 32.71% respectively. Recall value was improved by 13.69%, 43.95% and 34.59% through NB, RF and SVM. In terms of F1 score, NB increased it by 26.57%, RF increased the value by 40% and SVM improved it by 38.43%. Table III shows SVM classifier outperformed the other two classifiers in terms of Recall and F1 score. However, it yielded the worst performance on precision value. We also noticed that NB and RF showed similar results in HMIN II and NB outperformed RF model slightly. Figure 6 offers more obvious results comparison: After removing certain amount of noise that were identified by pure material feature from the network, substantial performance gains were achieved.

The third experiment was conducted through utilizing Feature III in the learning models. Table IV shows three classifiers performances obtained over HMIN II by using Feature III versus Feature I. From table IV we are able to observe the effectiveness of Feature III. Feature III enriches the meta path based-topological features by introducing the another measure, Hetesim, which is designed based on pair-wise random walk model, it evaluates how likely a pair of objects of different types on a meta path can meet at a certain point. Learning models leveraging Feature III demonstrated superior performance evaluated by recall and F1 score compared to using Feature I, even they performed slightly worse on precision. Figure 7 shows SVM is the optimal classifier by outperforming RF and NB, and RF and NB showed competitive predictive abilities.

For each group of experiment, we used 5-fold cross validation to report mean and standard deviation (number in parenthesis) values in terms of precision, recall and F1 and the optimal value for tuning learning model hyperparameters was identified by grid search.

There were several challenges while conducting experiments. First, the predictive performance was impacted by limited number of training samples obtained by manually labelling. The data size further shrunk after removing network noise. Our learning algorithms with different feature sets still achieved promising outcomes even so. Second, due to our specific application, only a single valid meta path exists between the publication and patent in the network, the number of meta paths limits the feature diversity. Furthermore, even we identified and removed errorneous eges including single-element composition, there were other kinds of noises caused by the heuristic approach to extract composition-patent connections still present in the network. For example, paths including compositions that are consist of multiple elements such as “Cu2O1”, “H1V1” were still labeled as negative samples in the training set.

Network	Method	Precision	Recall	F1
HMIN	NB	0.534(0.042)	0.628(0.081)	0.572(0.044)
	RF	0.560(0.069)	0.496(0.131)	0.518(0.104)
	SVM	0.532(0.028)	0.584(0.227)	0.536(0.129)
HMIN II	NB	0.738(0.029)	0.714(0.060)	0.724(0.040)
	RF	0.732(0.017)	0.714(0.081)	0.720(0.045)
	SVM	0.706(0.029)	0.786(0.055)	0.742(0.031)

TABLE III: Predictive Performance in HMIN and HMIN II by utilizing Feature I

V. Related Works

Use graph or network to model structured data, like relational database has been studied from different perspectives. Graph database or property graph store are commonly used models [21], [24], [25]. NoSQL graph databases, representative instance such as Neo4j [5] is developed which has an API

[5] https://neo4j.org/

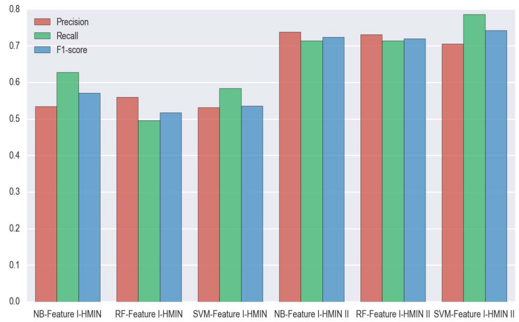


Fig. 6: Performance Comparison Between Baseline And Results From HMIN II By Leveraging Feature I

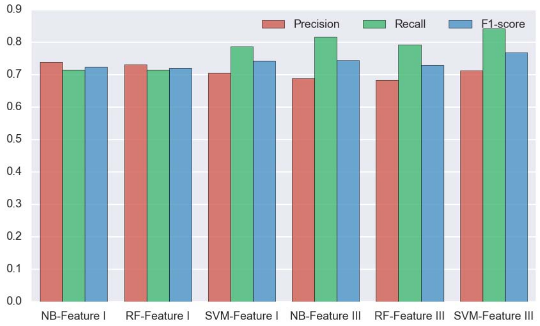


Fig. 7: Performance Comparison on Predictive Models by Using Feature I and Feature III over HMIN II

that is easy to use and supports efficient graph traversal. Graph databases usually encapsulate multiple attributes associated to an object within a single object node to save search time, therefore graph database is a node-centric solution. Another strategy is to translate relational data into RDF triple store, which is designed for storing object and relation in the form of a subject-predicate-object triple. Different from graph database, it is an edge-centric storage solution. Approaches to translate data from relational database to RDF base have been extensively studied. [26], [27] proposed to map database schema to RDF ontology, and attributes and attribute values of a relation become RDF predicates and literals respectively. Constructing a heterogeneous information network to represent data from relational database are studied in [28], [14], [29], [30]. Instead of being described as node-centric or edge-centric approach, path-centric is a proper description since meta-path is considered as an essential approach to explore the network. Distinct from these studies, our paper translated relational data to a heterogeneous information network and expanded it by integrated data from another source. Instead of solely considering topological features [11], [15], [13], we designed more types of features suitable to our needs to execute relevance search over the network.

Feature set	Method	Precision	Recall	F1
Feature I	NB	0.738(0.029)	0.714(0.060)	0.724(0.040)
	RF	0.732(0.017)	0.714(0.081)	0.720(0.045)
	SVM	0.706(0.029)	0.786(0.055)	0.742(0.031)
Feature III	NB	0.684(0.012)	0.792(0.082)	0.730(0.035)
	RF	0.712(0.015)	0.842(0.098)	0.768(0.044)
	SVM	0.688(0.012)	0.816(0.082)	0.744(0.035)

TABLE IV: Performance Comparison between Feature I and III on HMIN II

VI. Conclusion

In this work, we studied integrating structured data extracted from a materials science relational database with patent data from other source by constructing a heterogeneous materials information network, HMIN. We also investigated relevance between materials science literature and patents. Instead of performing relevance search on random publication-patent pairs, we explored paths between publication and patent nodes that follow certain meta path in order to capture path semantics and narrow the search space at the same time. Multiple sets of features were designed to measure objects connections from the perspectives of network topology and objects to fulfill various application purposes. We trained machine learning models leveraging those features to improve network data quality and explore publication-patent relevance. This is our initial effort to provide a heterogeneous information network prototype for integrating materials science data from multiple sources. Base on this infrastructure, we also designed useful features for network objects relevance search. Furthermore, even though the HMIN network is constructed to address the specific task proposed by materials researchers, HMIN-based approach can also be applied to other domains and various meta paths can be explored to fulfill people's requests. Future work could utilize advanced information extraction and entity recognition methods to retrieve network entities and remove network noise. Furthermore, current work only utilized one type of meta path to capture objects relatedness due to limited network information, we can enrich the current network by integrating other materials-related information to obtain richer types of meta paths to design more effective features and explore more meaningful applications for materials science community.

Acknowledgment

This research is a collaborate project between iCeNSA lab of University of Notre Dame and IBM Almaden Research Center. The author would like to thank Wolverton group in Northwestern University, they provided significant guidance and help regarding understanding Open Quantum Materials Database.

References

[1] S. R. Kalidindi and M. De Graef, "Materials data science: current status and future outlook," *Annual Review of Materials Research*, vol. 45, pp. 171–193, 2015.

[2] A. A. White, "Big data are shaping the future of materials science," *MRS Bull*, vol. 38, pp. 594–595, 2013.

[3] A. Agrawal and A. Choudhary, "Perspective: Materials informatics and big data: Realization of the "fourth paradigm" of science in materials science," *APL Materials*, vol. 4, no. 5, p. 053208, 2016.

[4] A. Dima, S. Bhaskarla, C. Becker, M. Brady, C. Campbell, P. Dessauw, R. Hanisch, U. Kattner, K. Kroenlein, M. Newrock *et al.*, "Informatics infrastructure for the materials genome initiative," *JOM*, vol. 68, no. 8, pp. 2053–2064, 2016.

[5] H. Wang and C. C. Aggarwal, "A survey of algorithms for keyword search on graph data," in *Managing and Mining Graph Data*. Springer, 2010, pp. 249–273.

[6] M. A. Rodriguez and P. Neubauer, "The graph traversal pattern," *arXiv preprint arXiv:1004.1001*, 2010.

[7] I. Robinson, J. Webber, and E. Eifrem, *Graph Databases: New Opportunities for Connected Data*. O'Reilly Media, Inc., 2015.

[8] D. Cai, Z. Shao, X. He, X. Yan, and J. Han, "Mining hidden community in heterogeneous social networks," in *Proceedings of the 3rd international workshop on Link discovery*. ACM, 2005, pp. 58–65.

[9] Y. Sun, J. Han, C. C. Aggarwal, and N. V. Chawla, "When will it happen?: relationship prediction in heterogeneous information networks," in *Proceedings of the fifth ACM international conference on Web search and data mining*. ACM, 2012, pp. 663–672.

[10] H. Fang, F. Wu, Z. Zhao, X. Duan, Y. Zhuang, and M. Ester, "Community-based question answering via heterogeneous social network learning," in *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*. AAAI Press, 2016, pp. 122–128.

[11] Y. Sun, J. Han, X. Yan, P. S. Yu, and T. Wu, "Pathsim: Meta path-based top-k similarity search in heterogeneous information networks," *Proceedings of the VLDB Endowment*, vol. 4, no. 11, pp. 992–1003, 2011.

[12] C. Meng, R. Cheng, S. Maniu, P. Senellart, and W. Zhang, "Discovering meta-paths in large heterogeneous information networks," in *Proceedings of the 24th International Conference on World Wide Web*. ACM, 2015, pp. 754–764.

[13] G. Fu, Y. Ding, A. Seal, B. Chen, Y. Sun, and E. Bolton, "Predicting drug target interactions using meta-path-based semantic network analysis," *BMC bioinformatics*, vol. 17, no. 1, p. 160, 2016.

[14] Y. Sun and J. Han, "Mining heterogeneous information networks: principles and methodologies," *Synthesis Lectures on Data Mining and Knowledge Discovery*, vol. 3, no. 2, pp. 1–159, 2012.

[15] Y. Sun, R. Barber, M. Gupta, C. C. Aggarwal, and J. Han, "Co-author relationship prediction in heterogeneous bibliographic networks," in *Advances in Social Networks Analysis and Mining (ASONAM), 2011 International Conference on*. IEEE, 2011, pp. 121–128.

[16] C. Shi, X. Kong, P. S. Yu, S. Xie, and B. Wu, "Relevance search in heterogeneous networks," in *Proceedings of the 15th International Conference on Extending Database Technology*. ACM, 2012, pp. 180–191.

[17] S. Kirklin, J. E. Saal, B. Meredig, A. Thompson, J. W. Doak, M. Aykol, S. Rühl, and C. Wolverton, "The open quantum materials database (oqmd): assessing the accuracy of dft formation energies," *NPJ Computational Materials*, vol. 1, p. 15010, 2015.

[18] D. Liben-Nowell and J. Kleinberg, "The link-prediction problem for social networks," in *CIKM '03 Proceedings of the twelfth international conference on Information and knowledge management*. ACM, 2003, pp. 1019–1031.

[19] Y. Yao and L. Holder, "Scalable svm-based classification in dynamic graphs," in *Data Mining (ICDM), 2014 IEEE International Conference on*. IEEE, 2014, pp. 650–659.

[20] L. Chi, B. Li, and X. Zhu, "Fast graph stream classification using discriminative clique hashing," in *Pacific-Asia Conference on Knowledge Discovery and Data Mining*. Springer, 2013, pp. 225–236.

[21] R. De Virgilio, A. Maccioni, and R. Torlone, "Converting relational to graph databases," in *First International Workshop on Graph Data Management Experiences and Systems*. ACM, 2013, p. 1.

[22] J. Leskovec and R. Sosič, "Snap: A general-purpose network analysis and graph-mining library," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 8, no. 1, p. 1, 2016.

[23] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "Smote: synthetic minority over-sampling technique," *Journal of artificial intelligence research*, vol. 16, pp. 321–357, 2002.

[24] W. Sun, A. Fokoue, K. Srinivas, A. Kementsietsidis, G. Hu, and G. Xie, "Sqlgraph: an efficient relational-based property graph store," in *Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data*. ACM, 2015, pp. 1887–1901.

[25] R. Angles and C. Gutierrez, "Survey of graph database models," *ACM Computing Surveys (CSUR)*, vol. 40, no. 1, p. 1, 2008.

[26] W. Hu and Y. Qu, "Discovering simple mappings between relational database schemas and ontologies," *The Semantic Web*, pp. 225–238, 2007.

[27] J. F. Sequeda, M. Arenas, and D. P. Miranker, "On directly mapping relational databases to rdf and owl," in *Proceedings of the 21st international conference on World Wide Web*. ACM, 2012, pp. 649–658.

[28] C. Shi, Y. Li, J. Zhang, Y. Sun, and S. Y. Philip, "A survey of heterogeneous information network analysis," *IEEE Transactions on Knowledge and Data Engineering*, vol. 29, no. 1, pp. 17–37, 2017.

[29] X. Yu, X. Ren, Y. Sun, B. Sturt, U. Khandelwal, Q. Gu, B. Norick, and J. Han, "Recommendation in heterogeneous information networks with implicit user feedback," in *Proceedings of the 7th ACM conference on Recommender systems*. ACM, 2013, pp. 347–350.

[30] C. Shi, C. Zhou, X. Kong, P. S. Yu, G. Liu, and B. Wang, "Heterecom: a semantic-based recommendation system in heterogeneous networks," in *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2012, pp. 1552–1555.