Intelligent Bibliometrics for Discovering the Associations between Genes and Diseases: Methodology and Case study

Mengjia Wu†
Australian Artificial Intelligence Institute
University of Technology Sydney
Sydney NSW Australia
Mengjia.Wu@student.uts.edu.au

Yi Zhang
Australian Artificial Intelligence Institute
University of Technology Sydney
Sydney NSW Australia
Yi.Zhang@uts.edu.au

ABSTRACT

Discovering disease-gene associations is an essential but challenging task in modern medicine. Within all the data-driven approaches targeting at this issue, literature-based knowledge discovery widely extends the discovering boundaries and uncovers implicit knowledge from unstructured textual data. However, most of the current literature-based methods require the involvement of specific expertise or prior knowledge. In this paper, we propose an adaptable and transferable methodology to 1) identify crucially genetic factors for a specific disease and 2) predict emerging genetic associations for the disease. Specifically, biomedical entities including diseases, chemicals, genes and genetic variations are extracted from literature data, then a heterogenous co-occurrence network is constructed and a semantic adjacency matrix is generated using the idea of Word2Vec. Following this, key genes and genetic variats are identified through centrality measurement on the network; emerging disease-gene associations are captured via a link prediction approach enhanced by the semantic matrix. We applied the proposed methodology to a literature dataset containing 54,219 scientific articles of atrial fibrillation (AF) to demonstrate its reliability. The results yielded a) crucial biomedical entities for AF highlighting five key gene groups and one potentially associated protein mutation; b) a list of emerging AF-genetic factors pairs that are worth in-depth exploration.

CCS CONCEPTS

• Network algorithms • Social and professional topics

KEYWORDS

Bibliometrics; Network analytics; Disease-Gene Association; Word embedding

1 Introduction

In modern medicine, deciphering disease-associated genes plays a vital role in the diagnosis, treatment and prevention of diseases. However, apart from the handful revealed molecular mechanisms and disease pathogenesis, there is still a substantial amount of disorders and abnormalities with their causes remaining underneath the tip of the iceberg, especially for the process and factors related to the inheritance and genetic basis. In the past few years, many efforts have been addressed on exploring the genetic basis of diseases. Although in the biomedical domain, genetic linkage analysis [1] and genome-wide association studies (GWAS) [2] are recognized as efficient and reliable methods in identifying disease-specific genes, the biggest challenge for those methods turns out to be the long list of gene candidates, resulting in the high economic costs, human efforts and trail risks for those experiments.

In the past decades, researchers established various medical ontologies and curated molecular networks to analyze and infer molecular interactions for diseases based on accumulated experimental and clinical experience [3-8]; Though these curated knowledge bases provide structuralized data sources for genetic discovery, their usages still face limitations from 1) the monotonicity of node category and the restriction of inference within the knowledge base framework; 2) the time lag of including novel discoveries and 3) the enormous cost from establishment and maintenance.

The explosively increasing biomedical literature and thriving text mining techniques provide a more open, real-time and economic pathway to solve those issues [9-11]. Most of the approaches using literature datasets still require certain pre-knowledge-based input for the target disease like its seed genes [12]. In this paper, we proposed an adaptable bibliometric methodology to infer disease-associated genetic factors by 1) excavating more categories (disease, chemical and four other genetic factors) of biomedical entity from the textual data; 2) utilizing the collected literature dataset to identify emerging genetic factors for the target

disease; 3) empowering our methodology purely data-driven and automatic without biomedical expertise or manual effort.

Our proposed methodology includes 1) a heterogenous bibliometric network [13]: the network is constructed based on biomedical literature with its nodes representing biomedical entities (e.g., diseases, chemicals, and genetic factors) and edges referring to the sentence-level co-occurrence between the connected nodes; 2) a Bioentity2Vec model: Using the idea of Word2Vec [14], all the biomedical entities are represented by computable vectors, from which an adjacency matrix containing their pairwise semantic similarities is generated; 3) network analytics: centrality measurement [15] is exploited to identify crucial diseases, chemicals and genetic factors within the network; a semantic similarity-enhanced link prediction approach [16] is proposed to improve the performance of predicting emerging disease-gene associations.

2 Methodology

The research framework of the proposed method is given in Figure 1.

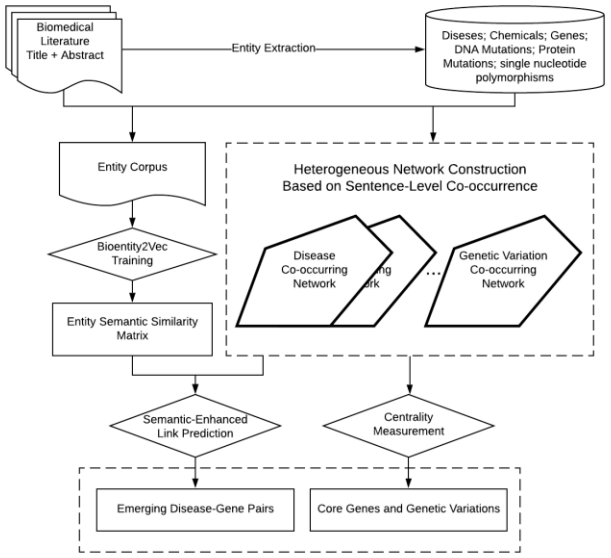

Figure 1: The Research Framework of the Heterogeneous Network Analytics Based Prediction Method

2.1 Entity Extraction and Network Construction

Four categories of biomedical entity are extracted from our literature dataset including: 1) disease: disease entity include disorders, symptoms, risk factors and complications related to the target disease; 2) chemical: chemical entity contains chemical elements, clinical drugs and other chemical compounds; 3) gene: the unit of hereditary information that occupies a fixed position (locus) on a chromosome; 4) Genetic Variant: genetic variants include DNA mutation, protein mutation and single nucleotide polymorphism (SNP): DNA mutation refers to the permanent change of a DNA sequence, protein mutation is the protein encoded by a gene with mutation and SNP [17] represents the normal variation of a single nucleotide in the gene sequence. Genes and genetic variants are the genetic factors we aim to predict in this study.

Working under the hypothesis that sentence-level occurrence indicates a strong association between entity pairs, we treated the extracted entities as nodes and assigned edges aligning with their sentence-level co-occurrence. In this way all the entities and their co-occurring relationships are transferred into a heterogeneous network, with its nodes representing entities and edges representing sentence-level co-occurrence as the following adjacency matrix explains:

$$A_{V_i^m V_j^n} = \begin{cases} CF(V_i^m, V_j^n) \ (if\ V_i^m\ and\ V_j^n\ co-occur\ in\ a\ sentence) \\ 0 \end{cases} \#(1)$$

where V_i^m represents the mth node in the ith category, $CF(V_i^m, V_j^n)$ refers to the record frequency of sentence-level co-occurrence of V_i^m and V_j^n;

The graph representation of the heterogeneous network is:

$$G = \left(V_K, E_{\frac{K(K+1)}{2}}\right) \#(2)$$

where V is the set of K categories of entity nodes and E is the set of $K(K+1)/2$ types of edges connecting different categories of nodes.

2.2 Bioentity2Vec Modelling

Enlightened by the idea of the Word2Vec model [14], we obtained the semantic similarities of biomedical entities from a context-based perspective. Word2vec is a well-accepted natural language model which can transfer word into vectors by projecting word its one-hot representation into a lower dimension and largely reserve their semantic meanings. In our case, biomedical entities are regarded as words and their consecutive sequences form our training corpus. We select the Skip-Gram as our training algorithm since it offer better fit on small datasets, the training process of Skip-Gram can be concluded as: given an entity $E(i)$ in a corpus, the probabilities of its nearby entities in a certain window size w will be calculated based on the probability of the given entity $E(i)$ [18], the global objective of is maximizing the following value which calculates the average conditional probability for all the windows within the corpus:

$$LF = \frac{1}{n}\sum_{i=1}^{n}\left(\sum_{-w\le j\le w, i\neq 0} \log_2 P\big(E(i+j)\big|E(i)\big)\right) \#(3)$$

Through the Bioentity2Vec training, each entity would be represented as a fixed dimensional vector; we could then calculate the pairwise similarity of entities via cosine similarity and generate an adjacency similarity matrix $S_{V_i^m V_j^n}$:

$$S_{V_i^m V_j^n} = \cos\left(v_{V_i^m}, v_{V_j^n}\right) = \frac{v_{V_i^m} \cdot v_{V_j^n}}{\sqrt{v_{V_i^m} \cdot v_{V_i^m}} \cdot \sqrt{v_{V_j^n} \cdot v_{V_j^n}}} \#(4)$$

where $v_{V_i^m}$ is the corresponding vector of entity node V_i^m.

2.3 Network Analytics

2.3.1 Centrality Measurement

Centralities are a set of measurements evaluating nodes' position and importance within the network [15]. In our study, degree centrality, closeness centrality and betweenness centrality are employed to identify the crucial nodes in our heterogeneous network; the three centralities respectively reflects the node's capacity of aggregating, disseminating and transferring information within the network, besides they have also been proved to be efficient in identifying key roles in biomedical entity networks [19]. The explanations of three centralities are as follows:

Degree Centrality (DC): the degree centrality measures the target node's direct influence to other nodes by calculating the proportion of degrees that the target node possesses. An entity with high degree centrality indicates that it directly interacts with a large number of other entities, the degree centrality is calculated as:

$$DC(V_i^m) = \frac{\sum_{j=1}^{K} \sum_{n=1}^{|V_j|} A_{V_i^m V_j^n}}{|V_K| - 1} \#(5)$$

where $|V_K|$ is the number of all K categories of nodes in the network and $|V_j|$ represents the node number in the j th category;

Closeness Centrality (CC): this closeness centrality calculates the target node's topological distance to all the other nodes in the network, the higher closeness centrality indicates the entity's stronger capacity to reach all the other nodes within the network, the closeness centrality is calculated as:

$$CC(V_i^m) = \frac{|V_K| - 1}{\sum_{j=1}^{K} \sum_{n=1}^{|V_j|} d_{V_i^m V_j^n}} \#(6)$$

where $d_{V_i^m V_j^n}$ is the topological distance from node V_i^m to node V_j^n;

Betweenness Centrality (BC): the betweenness centrality of a target node is the ratio of shortest paths between other node pairs that pass through the target node, it indicates the node's potential to bridge other nodes in the network. In our network, a higher betweenness centrality reflects that the entity is highly likely to be an important connector or transmitter:

$$BC(V_i^m) = \frac{2\sum_{x,y=1}^{K} \sum_{a=1}^{|V_x|} \sum_{b=1}^{|V_y|} \frac{\sigma\left(V_x^a V_y^b\right)_{V_i^m}}{\sigma\left(V_x^a V_y^b\right)}}{(|V_K| - 1)(|V_K| - 2)} \left(V_i^m \neq V_x^a \neq V_y^b\right) \#(7)$$

where $\sigma(V_x^a V_y^b)$ is the number of all the shortest paths from node V_x^a to V_y^b and $\sigma(V_x^a V_y^b)_{V_i^m}$ is the number of paths that pass through node V_i^m among all of them.

To further comprehensively measure the entities' importance using the three centralities, non-dominated sorting is used to combine the three centrality rankings for each entity category. Non-dominated sorting is a multiple-objective optimization method which re-rank the multi-dimensional scalable individuals by dominating relationships of one individual against another, after non-dominated sorting the individuals will be divided into several consecutive Pareto fronts according to their domination counts [20]. The pseudo-code of non-dominated sorting is presented in Figure 2.

```
for V_i in V_K:
    for node V_i^m in V_i:
        Domination[V_i^m] = 0      # Initialize the Domination Counts
        for node V_i^n in category i (m ≠ n):
            if
        DC(V_i^m) ≥ DC(V_i^n) and CC(V_i^m) ≥ CC(V_i^n) and BC(V_i^m)
                                ≥ BC(V_i^n) and
        (DC(V_i^m) = DC(V_i^n) and DC(V_i^m) = DC(V_i^n) and DC(V_i^m)
                                = DC(V_i^n)) == False:
            Domination[V_i^m] += 1
    Resort all the nodes in V_i by descending Domination counts
Output: K Sorting results for K categories
```

Figure 2: The pseudo-code for non-dominated sorting

2.3.2 Semantic Similarity-Enhanced Link Prediction

Link prediction is an approach used to fulfil the incomplete network or in predicting future emerging links in networks [21]. Based on our previous pilot studies, resource allocation (RA) is the best-performing algorithm among all neighbor-based comparisons [16], it assumes that every single node in a network has a unit of resources and the common neighbor of two nodes plays the role of the transmitter, evenly distributing the unit of resource to its connected nodes. The resource allocation index of any two unconnected nodes is the summarization of those resources allocated by all their common neighbors, which indicates the potential link strength between the two nodes. The higher this index value is, the greater is the possibility of a link emerging between them.

A weighted version of this algorithm was also introduced for applications in weighted networks [22], assuming that the co-occurring frequency and semantic similarity of connected node pairs could enhance the predicting accuracy, we incorporated the cosine similarities matrix generated from the Bioentity2Vec model before to modify this algorithm, the modified resource allocation index is calculated as:

$$R_{V_i^m V_j^n} = \sum_{V^t \in \Gamma(V_i^m) \cap \Gamma\left(V_j^n\right)} \frac{CF(V_i^m, V^t)\left|S_{V_i^m V^t}\right| + CF\left(V^t, V_j^n\right)\left|S_{V^t V_j^n}\right|}{\sum_{V^k \in \Gamma(V^t)} CF(V^k, V^t) S\left(S_{V^k, V^t}\right)} \quad \#(8)$$

where $\Gamma(V_i^m)$ denotes the set of neighbor nodes of V_i^m and $CF(V_i^m, V^t)$ is the co-occurring frequency of V_i^m and V^t, $S_{V_i^m V^t}$ denotes the semantic similarity of V_i^m and V^t.

Using the genetic factors that haven't co-occurred with the target disease before as our input, we could generate another final output through our link prediction approach: a ranking list of genetic factors with their corresponding predicting scores. The highly ranked ones are worth being validated by further biomedical experiments.

3 Case Study

Atrial fibrillation (AF) is the most common form of cardiac arrhythmia. The progress of AF is closely related to atrial size and the extent of atrial fibrosis, both of which are affected by genetic factors. Though several gene groups and mutations have been linked to AF, clinical evidence and mechanistic explanations are still far from being enough to integrate the knowledge of genetic risk factors into clinical practice [23]. Ongoing research is investigating discovered genes and seeking new gene associations. For these reasons, the choice of atrial fibrillation as our research topic here is both an appropriate and worthwhile undertaking.

3.1 Data Collection

PubMed is a biomedical literature search engine which comprises more than 30 million citations from MEDLINE database, PMC citations and other online book resources. We used "atrial fibrillation" as the searching term in PubMed and refined the search results by restricting the fields "species" to "humans", MeSH searching was adopted to promise precise AF-related search results and no restriction was applied to publication date. In all, 54,219 records were retrieved from the exact searching query:

"("Atrial Fibrillation"[Mesh] AND Humans[Mesh])"

Search Date: 2020/04/28

3.2 Entity Extraction and Network Construction

We exploited the Pubtator [24], Medical Subject Headings (MeSH) [1], NCBI Homo-Sapiens Gene Dictionary[2] and dbSNP database to extract biomedical entities from the collected literature dataset. Pubtator is a deep learning-based entity extraction tool developed by the National Library of Medicine (NLM), it can automatically extract categorized biomedical concepts from the free texts. MeSH is a medical thesaurus covering all the disease and chemical concepts, NCBI Homo-Sapiens Gene Dictionary is a gene dictionary of homo-sapiens species; dbSNP database embodies genetic variants within the human genome, the most of discovered DNA mutations, protein mutation and SNP can be matched to a specific SNP ID. Generally, concepts cannot be mapped to these dictionaries would be excluded.

Using Pubtator API and we extracted 577,809 raw biomedical concepts from the 54,219 records. With the aid of MeSH and NCBI gene dictionary, we cleaned noise concepts and consolidated all the synonyms, generating 6,318 identical biomedical entities; furtherly we excluded those entities that never co-occurred with others before (i.e., the isolated nodes that are not connected to any other nodes in the network) and ended up with 5,838 nodes. The stepwise results are given in Table 1.

Table 1: Stepwise Pre-processing Steps for Entity Extraction

Category	Raw Concepts	Description	Nodes
Disease	440,610	Remove those diseases and chemicals that cannot map to MeSH, like "cardioembolic", "JAGS", "nonvitamin", etc; Consolidation Based on MeSH	2,040
Chemical	104,072		2,004
Gene	31,209	Exclude genes that do not belong to homo-sapiens species; Consolidate gene aliases based on NCBI Homo-Sapiens Gene Dictionary	1,413
Genetic Variant	223 #1 770 #2 925 #3	Remove genetic variants without a clear varying locus (cannot match to a SNP ID) and match the valid ones to their SNP IDs	381
Total	577,809		5,838

Note: #1 DNA mutation; #2 Protein mutation; #3 SNP.

Through the sentence-level co-occurrence analysis, we constructed the heterogeneous network with 5,838 nodes and 48,988 edges.

[1] More information could be found at https://www.ncbi.nlm.nih.gov/mesh/
[2] More information could be found at https://www.ncbi.nlm.nih.gov/gene/

3.3 Identification of Core Entities for AF

To identify the crucial biomedical entities in the AF progress, we respectively calculated the degree, closeness and betweenness centralities for all the nodes in our heterogeneous network, the gene entities in the top 20 of each centrality ranking list are given in Table 2.

Table 2: Top 20 Genes with Respectively the Highest Degree, Closeness and Betweenness Centrality

	Degree Centrality	Closeness Centrality	Betweenness Centrality
1	CRP	CRP	KCNA5
2	IL6	NPPB	CRP
3	NPPB	ACE	SCN5A
4	F2	F2	F2
5	ACE	IL6	ACE
6	AGT	INS	AGT
7	F10	COX8A	TBX5
8	KCNA5	BID	PITX2
9	SCN5A	VWF	IL6
10	FGB	F10	SOX5
11	INS	AGT	KCNQ1
12	COX8A	REN	TRPM7
13	TGFB1	FGB	TRPC3
14	MMP9	NPPA	GJA5
15	TNF	CD59	HCN4
16	VWF	MMP9	F10
17	GJA5	TNF	TNNI3
18	PITX2	SELP	TRPM6
19	SELP	TGFB1	TRPM4
20	REN	TNNI3	GATA6

From the observation of their centrality characteristics, we classified the gene nodes into 5 groups and analyzed their topological features and node composition with the aid of NCBI gene database[3] and biomedical literature investigation:

1. High Degree & Closeness & Betweenness Centralities: These topological features reflect the dominating positions of these nodes in the AF heterogeneous network. *CRP*, *IL6*, *AGT*, *ACE*, *F2* and *F10* are genes belong to this group; they are all early-discovered genes and have broad functioning ranges and massive interactions with other entities. Specifically, C-reactive protein (*CRP*) and interleukin 6 (*IL6*) are genes that function in inflammation reaction and the immune-related activities, with their encoding product's levels associated with the prediction of a wide variety of cardiovascular events including AF; Angiotensin I converting enzyme (*ACE*) and angiotensinogen (*AGT*) are two chain functioning genes with the former encoding pre-angiotensinogen which would be cleaved by the angiotensin I converting enzyme encoded by the later, the product angiotensin II from this process is a significant protein in controlling the blood pressure (BP) and fluid-electrolyte balance, so both BP related symptoms like hypertension and electrolyte adjustment chemicals are associated with the two genes; coagulation factor II (*F2*) and coagulation factor X (*F10*) are genes that encode major coagulation factors to intermediate blood clotting and hemorrhagic conditions related to AF. Conclusively, activities of genes in this group may not directly result in AF but their functions engage the most primary and foundational molecular mechanisms in the progress of AF.
2. High Degree & Betweenness Centralities but low Close Centrality: From a graph theory perspective, we can interpret this group of gene nodes as the critical but localized controllers. This group of genes includes *KCNA5* and *SCN5A*, both of which are also seed genes for AF with their associations with AF already revealed: *KCNA5* and *SCN5A* respectively encode proteins for potassium and sodium voltage-gated channels, the loss or alternation of those channels' function have a direct influence on the action potential and electrical activity of cardiomyocytes which may further lead to AF. For this group, gene nodes bridle the ion channel-related entities including symptoms and chemicals but one node majorly only covers one certain type of ion channel.
3. High Degree & Close Centralities but low Betweenness Centrality: This group of gene nodes are inclined to be central nodes in their sub-components with high independence. *NPPB*, *FGB*, *COX8A*, *VWF* and *INS* are genes in this group, in which *NPPB* encodes the cardiac hormone with its blood concentration indicating the heart failure; *FGB* encodes the beta component of fibrinogen with whose deficiency or mutation leading to afibrinogenemia; COX8A encodes the terminal enzyme of the respiratory chain related to ATP synthesis and cardiomyopathy; *VWF* encodes a glycoprotein involved in hemostasis; *INS* is the gene in charge of insulin synthesis which is the critical chemical in diabetes. We could conclude that genes in this group are most directly associated with the most common AF risk factors or complications rather than AF itself.
4. High Betweenness Centrality only: This group of genes including *TBX5*, *SOX5* and *PITX2* are less correlated to AF

[3] More information could be found at https://www.ncbi.nlm.nih.gov/gene/

compare with the aforementioned ones but the high betweenness centrality indicates their potential to connect AF with other entities. For example, *TBX5* encodes transcription factor that is associated with heart developmental process and its mutation may result in a heart-affecting developmental disorder named Holt-Oram syndrome.

5. High Closeness Centrality only: Topologically this group of gene it's not the core nodes but still globally associated with the other AF entities. *BID* is the only gene in this group and it regulates cell's apoptosis which is not a particular biological process for AF but generally correlated with other entities.

Moreover, with applying non dominated sorting to all the node categories, we re-ranked the nodes for each category and generated Table 3 to further identify the other crucial entities.

Table 3: Key Entities Identified from Centrality Measurements

	Top 20 Results by Non-dominating Sorting
Disease	Atrial Fibrillation; Stroke; Heart Failure; Hypertension; Hemorrhage; Diabetes Mellitus; Fibrosis; Myocardial Infarction; Cerebral Infarction; Ischemia; Thromboembolism; Death; Thrombosis; Inflammation; Coronary Artery Disease; Tachycardia; Ventricular Fibrillation; Tachycardia, Supraventricular; Neoplasms; Atrioventricular Block
Chemical	Warfarin; Calcium; Amiodarone; Potassium; Digoxin; Ethanol; Verapamil; Sodium; Oxygen; Quinidine; Aspirin; Vitamin K; Glucose; Cholesterol; apixaban; Sotalol; Nitrogen; Magnesium; Heparin; Propafenone
Gene	CRP; F2; ACE; IL6; AGT; F10; SCN5A; NPPB; KCNA5; PITX2; FGB; GJA5; TNNI3; INS; TNF; TGFB1; VWF; KCNQ1; SERPINE1; AGTR1
SNP	rs2200733; rs6795970; rs2106261; rs2108622; rs3789678; rs13376333; rs17042171; rs1805127; rs7539020; rs11568023; rs10033464; rs3807989; rs7193343; rs3918242; rs3825214; rs16899974; rs699; rs7164883; rs6584555; rs10824026

Apart from the genes which were evaluated before, we examined the reliabilities of other core entities respectively by looking through ClinVar [25] and SNPedia [26]; literature investigation was still used to provide supplementary evidence.

The disease list presents the most common risk factors (*Hypertension* and *Diabetes Mellitus*), symptoms (*Inflammation* and *Thrombosis*), complications (*Stroke* and *Heart Failure*) and other correlating diseases (*Hemorrhage*, *Fibrosis* and *Cerebral Infarction* etc.) of AF, which are frequently reported in clinical cases. The chemical list highlights the regular treating drugs (*Warfarin, Amiodarone, Digoxin, Verapamil, Quinidine, Aspirin, Apixaban, Sotalol, Heparin* and *Propafenone*), the known pathological molecular mechanisms (*Calcium, Sodium, Potassium* and *Magnesium* channels) and risk factors (*Ethanol* and *Glucose*) of AF; Nitrogen and Oxygen are two other leading elements, resulting from the role of blood oxygen concentration as an indicating index and NOx as a risk factor in the progress of AF.

To sum up, we regard the whole centrality measurement and non-dominated sorting as a carding process for the AF-related biomedical entities, through which we not only captured the comprehensively critical biomedical entities but also gained clues on some potential AF-associated entities.

3.4 Link Prediction Validation

Before implementing the modified link prediction to our heterogeneous network, we performed a validating experiment on the rolled-back data to verify our algorithm's usefulness on disease-gene association prediction. Two other link prediction algorithms were selected as our baselines:

1. The original resource allocation (RA): the original version of resource allocation index mentioned before;
2. the co-occurring frequency (CF) weighted version of resource allocation: this version of RA adopted the same assumption with RA but uses weight ratio instead of degree proportion to calculate the resource diffusion, in our study edge weight is by the entity's co-occurring frequency;

The validation experiment was designed as follows: We rolled back our dataset by a five-year gap and constructed a corresponding network (i.e., rolled-back network), and the newly researched k AF-linked genes or SNPs in the latest five years were collected as true labels. Then, two baselines, as well as our modified version, were applied to the rolled-back network to predict emerging links. The predictive results are a mixed ranking list of genes and SNPs with their corresponding predictive scores. If a gene or SNP in true labels was correctly predicted in the top n (n is a selectable threshold according to predictive requirements, initially it was set as k) predicting list, it would count a true positive (TP), otherwise, it would constitute a false negative (FN). The outcomes are provided in Table 4.

$$Recall = \frac{TP}{TP + FN}$$

Table 4: Recall Results for Validating Experiment

Index	Algorithms		
	Unweighted	***Weighted***	***Modified (Proposed)***
Top Recall	0.245	0.212	**0.283**
Top 100 Recall	0.436	0.392	**0.502**
Top 200 Recall	0.610	0.632	**0.742**

From the results we can see that the modified link prediction approach method beat the other two baselines. This experiment fully proved the efficiency of our proposed method. Briefly, in the top 200 list, our algorithm successfully captured 74% of the correct genetic factors which would appear in the following five years, from an applying standpoint it largely reduces the necessary human workload by using the predictive shortlist to select candidate genetic factors.

3.5 AF-related Genetic Factors Prediction

We applied our proposed method to the heterogeneous network and obtained a list of disease-gene pairs listed in Table 5. Then we empirically searched evidence from literature to validate our predication and found that the top 10 could all be linked according to literature review. The predicted SNPs are attracting more attention in the latest research of AF [27] and the other genes are also more frequently studied in the given literature. For example, rs337711 (ranked 2 in our list), is a genetic variation in KCNN2 engaging with the potassium voltage-gated channels and has influence on the action potential and electrical activity of cardiomyocytes related to AF [23]. The association of rs11264280 (ranked 3 in our list) is reported to be contract from two separate experiments [27, 28]. One of PKP2 (ranked 5 in our list) mutations was reported to has a potential influence on atrial size and another deletion mutation was reported to be related to the occurrence of lone AF [29, 30]. Warfarin, a commonly prescribed anticoagulant for nonvalvular AF, is facing a medication shift because of the adverse effect of valvular calcification due to gene MGP (ranked 9 in our list) [31]. The combination of TFF3 (ranked 10 in our list) and P3NP has the potential to be a biomarker of atrial fibrillation [32].

Table 5: Predicting Results for AF-related Genetic Factors

	Candidate genetic factors	Predicting Score	Literature Evidence
1	Gene | BGLAP	3.13	[33]
2	SNP | rs337711	2.13	[34, 35]
3	SNP | rs11264280	2.13	[27, 28]
4	Gene | HP	2.11	[36]
5	Gene | PKP2	2.05	[29, 30]
6	Gene | DUOX2	1.91	[37]
7	Gene | OLR1	1.88	[38]
8	Gene | VIM	1.87	[39]
9	Gene | MGP	1.84	[31]
10	Gene | TFF3	1.83	[32, 40]

4 Discussion

Identifying disease-gene associations is a critical part of modern medicine. Increasing evidence reveals the strong links between genetics and human health, predicting those associations that will provide effective decision support for medical and clinical researchers. Compared to the earlier published version, we involved more categories of biomedical entities and centrality measurement approach to modify the research framework, through which we respectively added genetic variants into our predicting scope and identified critical genetic factors that function in the pathogenesis of the target disease.

This paper proposed a hybrid method for predicting the associations between diseases and genes from the standpoint of bibliometrics, based on the co-occurrence and semantic similarities of genetic factors and diseases. The advantages of our method are that we made use of text data in the latest published papers and took semantic relationships into consideration. By designing a cross-validation experiment, we compared our hybrid method with other classical ones and found that our method shows better performance. Furthermore, the predicted relationships are identified in the latest studies, proving the credibility and validity of our proposed method.

Apart from this case study on gene-related atrial fibrillation, the proposed method is expected to be applied to a broad range of investigations on discovering the relationships between genes and specific diseases. Such efforts could be expected to provide extensive and objective insights from global scientific articles to support decision making in related medical research and clinical practices, such as helping researchers identify unknown relationships between specific genes and diseases and propose effective treatments from globally published research, studies, and cases.

There are also several limitations of this paper that may require further investigation in future studies: 1) Although we exploited the developed Pubtator as our biomedical entity extraction tool, there exist inevitably some false positives in the entity extraction process beyond our control, how to avoid the impact of using those toolkits is one of our concerns in the future; 2) we emphasized on the completeness of collecting strong association by adopting co-occurrence analysis, but this process would include some negative associations like “A is not associated with B”, in the future studies, sentiment analysis would be a promising approach to reduce the false positives. 3) Limited expertise is employed here to validate and explain our predicting results. In the future study, we plan to establish a cardiovascular specialist panel to provide expert instructions and interpretations of our outcomes.

ACKNOWLEDGMENTS

An early version of this work has been published in the Proceedings of the 2020/2021 Portland International Center for Management of Engineering and Technology. This work is supported by the Australian Research Council under Discovery Early Career Researcher Award DE190100994.

REFERENCES

[1] J. Ott, Analysis of human genetic linkage. JHU Press, 1999.
[2] W. S. Bush and J. H. Moore, "Genome-wide association studies," PLoS computational biology, vol. 8, no. 12, 2012.
[3] A.-L. Barabási, N. Gulbahce, and J. Loscalzo, "Network medicine: a network-based approach to human disease," Nature reviews genetics, vol. 12, no. 1, pp. 56-68, 2011.
[4] J.-F. Rual et al., "Towards a proteome-scale map of the human protein–protein interaction network," Nature, vol. 437, no. 7062, pp. 1173-1178, 2005.
[5] K. Venkatesan et al., "An empirical framework for binary interactome mapping," Nature methods, vol. 6, no. 1, p. 83, 2009.
[6] D. Szklarczyk et al., "STRING v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets," Nucleic acids research, vol. 47, no. D1, pp. D607-D613, 2019.
[7] S. van Dam, U. Vosa, A. van der Graaf, L. Franke, and J. P. de Magalhaes, "Gene co-expression analysis for functional classification and gene–disease predictions," Briefings in bioinformatics, vol. 19, no. 4, pp. 575-592, 2018.
[8] E. Anastasiadou, L. S. Jacob, and F. J. Slack, "Non-coding RNA networks in cancer," Nature Reviews Cancer, vol. 18, no. 1, p. 5, 2018.
[9] C. M. Friedrich, H. Dach, T. Gattermayer, G. Engelbrecht, S. Benkner, and M. Hofmann-Apitius, "for Biomedical Knowledge Discovery," Global Healthgrid: E-Science Meets Biomedical Informatics: Proceedings of HealthGrid 2008, vol. 138, p. 165, 2008.
[10] S. Henry and B. T. McInnes, "Literature based discovery: models, methods, and trends," Journal of biomedical informatics, vol. 74, pp. 20-32, 2017.
[11] G. E. Heo, Q. Xie, M. Song, and J.-H. Lee, "Combining entity co-occurrence with specialized word embeddings to measure entity relation in Alzheimer’s disease," BMC Medical Informatics and Decision Making, vol. 19, no. 5, p. 240, 2019.
[12] A. Özgür, T. Vu, G. Erkan, and D. R. Radev, "Identifying gene-disease associations using centrality on a literature mined gene-interaction network," Bioinformatics, vol. 24, no. 13, pp. i277-i285, 2008.
[13] M. Callon, J.-P. Courtial, W. A. Turner, and S. Bauin, "From translations to problematic networks: An introduction to co-word analysis," Information (International Social Science Council), vol. 22, no. 2, pp. 191-235, 1983.
[14] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," arXiv preprint arXiv:1301.3781, 2013.
[15] L. C. Freeman, D. Roeder, and R. R. Mulholland, "Centrality in social networks: II. Experimental results," Social networks, vol. 2, no. 2, pp. 119-141, 1979.
[16] T. Zhou, L. Lü, and Y.-C. Zhang, "Predicting missing links via local information," The European Physical Journal B, vol. 71, no. 4, pp. 623-630, 2009.
[17] T. D. Arias, L. Jorge, and R. Barrantes, "Uses and misuses of definitions of genetic polymorphism. A perspective from population pharmacogenetics," British journal of clinical pharmacology, vol. 31, no. 1, p. 117, 1991.
[18] X. Rong, "word2vec parameter learning explained," arXiv preprint arXiv:1411.2738, 2014.
[19] A. Al-Aamri, K. Taha, Y. Al-Hammadi, M. Maalouf, and D. Homouz, "Analyzing a co-occurrence gene-interaction network to identify disease-gene association," BMC bioinformatics, vol. 20, no. 1, p. 70, 2019.
[20] Y. Yuan, H. Xu, and B. Wang, "An improved NSGA-III procedure for evolutionary many-objective optimization," in Proceedings of the 2014 Annual Conference on Genetic and Evolutionary Computation, 2014, pp. 661-668.
[21] D. Liben-Nowell and J. Kleinberg, "The link-prediction problem for social networks," Journal of the American society for information science and technology, vol. 58, no. 7, pp. 1019-1031, 2007.
[22] L. Lü and T. Zhou, "Link prediction in weighted networks: The role of weak ties," EPL (Europhysics Letters), vol. 89, no. 1, p. 18001, 2010.
[23] J. Feghaly, P. Zakka, B. London, C. A. MacRae, and M. M. Refaat, "Genetics of atrial fibrillation," Journal of the American Heart Association, vol. 7, no. 20, p. e009884, 2018.
[24] C.-H. Wei, A. Allot, R. Leaman, and Z. Lu, "PubTator central: automated concept annotation for biomedical full text articles," Nucleic acids research, vol. 47, no. W1, pp. W587-W593, 2019.
[25] M. J. Landrum et al., "ClinVar: public archive of interpretations of clinically relevant variants," Nucleic acids research, vol. 44, no. D1, pp. D862-D868, 2016.
[26] M. Cariaso and G. Lennon, "SNPedia: a wiki supporting personal genome annotation, interpretation and analysis," Nucleic acids research, vol. 40, no. D1, pp. D1308-D1312, 2012.
[27] Y. Pan, Y. Wang, and Y. Wang, "Investigation of Causal Effect of Atrial Fibrillation on Alzheimer Disease: A Mendelian Randomization Study," Journal of the American Heart Association, vol. 9, no. 2, p. e014889, 2020.
[28] X. Wang et al., "Rs17042171 at chromosome 4q25 is associated with atrial fibrillation in the Chinese Han population from the central plains," Journal of Central South University. Medical sciences, vol. 43, no. 6, p. 594, 2018.
[29] C. Yeung, A. Enriquez, L. Suarez-Fuster, and A. Baranchuk, "Atrial fibrillation in patients with inherited cardiomyopathies," Ep Europace, vol. 21, no. 1, pp. 22-32, 2019.
[30] S. Alhassani, B. Deif, S. Conacher, K. S. Cunningham, and J. D. Roberts, "A large familial pathogenic Plakophilin-2 gene (PKP2) deletion manifesting with sudden cardiac death and lone atrial fibrillation: Evidence for alternating atrial and ventricular phenotypes," HeartRhythm case reports, vol. 4, no. 10, pp. 486-489, 2018.
[31] R. F. Reilly and N. Jain, "Warfarin in nonvalvular atrial fibrillation—Time for a change?," in Seminars in dialysis, 2019, vol. 32, no. 6: Wiley Online Library, pp. 520-526.
[32] I. P. Doulamis et al., "Proteomic profile of patients with atrial fibrillation undergoing cardiac surgery," Interactive CardioVascular and Thoracic Surgery, vol. 28, no. 1, pp. 94-101, 2019.
[33] S. A. Millar, H. Patel, S. I. Anderson, T. J. England, and S. E. O’Sullivan, "Osteocalcin, vascular calcification, and atherosclerosis: a systematic review and meta-analysis," Frontiers in Endocrinology, vol. 8, p. 183, 2017.
[34] S. A. Lubitz, B. A. Yi, and P. T. Ellinor, "Genetics of atrial fibrillation," Cardiology clinics, vol. 27, no. 1, pp. 25-33, 2009.
[35] X. Wang et al., "Rs17042171 at chromosome 4q25 is associated with atrial fibrillation in the Chinese Han population from the central plains," Zhong nan da xue xue bao. Yi xue ban= Journal of Central South University. Medical sciences, vol. 43, no. 6, p. 594, 2018.
[36] S. Mohanty et al., "P2663Increase in haptoglobin level at the immediate post-ablation period predicts ablation success in patients with atrial fibrillation: Results from a prospective study (IMPACT II)," European Heart Journal, vol. 38, no. suppl_1, 2017.
[37] M. M. Kizys et al., "DUOX2 mutations are associated with congenital hypothyroidism with ectopic thyroid gland," The Journal of Clinical Endocrinology & Metabolism, vol. 102, no. 11, pp. 4060-4071, 2017.
[38] T. Skarpengland et al., "Increased levels of lectin-like oxidized low-density lipoprotein receptor-1 in ischemic stroke and transient ischemic attack," Journal of the American Heart Association, vol. 7, no. 2, p. e006479, 2018.
[39] S. Kwon et al., "Fluctuating renal function and the risk of incident atrial fibrillation: a nationwide population-based study," Scientific Reports, vol. 9, no. 1, pp. 1-8, 2019.
[40] M. Brankovic et al., "Utility of temporal profiles of new cardio-renal and pulmonary candidate biomarkers in chronic heart failure," International journal of cardiology, vol. 276, pp. 157-165, 2019.