Disease Gene Knowledge Map Construction Technology

Yayun Shu, Jiamei Liu, Shanshan Xiong

Software College, Jilin University, Changchun, Jilin, China

Abstract. Based on the research of domestic and international generic and bio-industry-specific knowledge maps, we use the fusion of multiple disease-related gene banks (structured data) to explore the automated construction methods and standardization processes of disease genes and explore the relevance of pathogenic gene knowledge maps. Applications include knowledge-based gene search and prediction of pathogenic genes.

1.Introduction

Since the 1980s, along with the launch of the Human Genome Project (HGP) (Heilig et al., 2004), the new discipline of bioinformatics as a combination of biology, computer science and life science has continued to flourish. It mainly studies biological data information and its interactions and laws, which is an important part of the rapid development of biology as a whole. The application of bioinformatics can greatly increase the efficiency of biological and medical research, shorten the research cycle. However, with the advent of the age of big data, there has also been explosive growth in biological data and this has led to a series of problems: On the one hand, increasing data sets unprecedented demands on the collection and processing of information; On the other hand, how to predict unknown relationships between entities (such as identifying protein genes and interpreting the genetic code of genes, etc.) based on known information and knowledge is a huge challenge currently facing them (Jiang Xin, 2005).

The knowledge base organizes human knowledge into structured knowledge systems and describes the relationships between entities in the real world. The main research objectives of the knowledge base are: to extract structured knowledge from semi-structured Internet information, and to automatically integrate related applications such as knowledge base construction and service knowledge reasoning. Knowledge representation is the basis for knowledge acquisition and application (relationship prediction, etc.) (Liu Zhiyuan, Sun Maosong, Lin Yankai, & Xie Ruobing, 2016). Similar to Knowledge Discovery in Database (KDD) (Wang Zhihong, 2006), a reasonable knowledge representation can construct global-based semantic information for different entities and relationships in a large number of entities and relationships. Compared with traditional logic-based knowledge representation methods, statistical learning adapts more to the current knowledge base and has better computational efficiency.

Building disease gene banks can extract known disease gene entities and relationships, internet knowledge use knowledge representation methods to discover hidden relationships and laws.

We will extract entities and relationships from multiple existing gene ontology such as Gene Ontology (GO) and HPO, construct an ontology-based disease gene knowledge base. And based on classical models in this field (such as TransE model) (Bordes, Usunier, Garcia-Duran, Weston, & Yakhnenko, 2013), a better knowledge representation learning model is proposed, knowledge inference is performed, and the entity in the knowledge base is effectively mitigated. The relationship obeys the data sparse problem caused by the long-tailed distribution. Based on the biological data set,

the implicit relations between various biological entities are discovered. Analyzing the relationship between these relationships provides a reliable basis for biologists. It can also provide a novel idea for their research and is conducive to the development of bioinformatics.

2. Typical knowledge map

2.1 Freebase

Freebase is a large collaborative knowledge base composed of metadata, the content of which is mainly contributed by the contributions of its community members. It integrates many online resources, including content from some private wiki sites. Freebase is committed to creating a library that allows everyone quick access. It was developed by Metaweb, a US software company, and was openly operational in March 2007. July 16, 2010 was acquired by Google. On December 16, 2014, Google announced that it will close Freebase after six months and migrate all data to Wikidata.

2.2 DBpedia

DBpedia is a cross-language comprehensive database developed for the LOD project. The basic idea is to extract the existing structured knowledge from the wiki and store it in PDF format. Based on this, query and apply it. So far, DBpedia contains over 3 billion PDF tuples. In addition, YAGO also extracts knowledge from the wiki and also takes into account WordNet's semantic information to build richer entity relationships. YAGO contains over 10 million entities and 1.2 pieces of knowledge, and a detailed breakdown of these entities and relationships.

From the above knowledge map, it can be seen that the construction and embedding representation of the open domain knowledge map has been booming. At the same time, specific domain knowledge maps, especially genetic knowledge maps, are still rare. We explore the database of multiple disease-related genes and build knowledge maps of disease genes using ontology-based knowledge mapping methods. We also use the TransH model (Wang, Zhang, Feng, & Chen, 2014) to reason about knowledge and use the knowledge map to predict possible pathogenic genes to provide biologists with reliable data.

2.3 Related work

This article mainly extracts related entities and relationships from multiple existing knowledge bases such as OMIM and HPO and builds an ontology-based disease gene knowledge base. And use multiple optimization translation models for knowledge reasoning to discover the implicit relationships between various biological entities. Analyzing the relationship between these relationships provides a reliable basis for biologists. It can also provide a novel idea for their research and is conducive to the development of bioinformatics.

3. Knowledge base construction

3.1 Overview

Knowledge map is a networked knowledge system constructed with the semantic network as the skeleton. It can capture and present the semantic relations between domain concepts and make the trivial and scattered knowledge on the internet connect with each other. Knowledge map of disease genes is mainly based on the genetic map system constructed with mutation-causing genes as the skeleton, the existing database resources are the filling content of the knowledge map.

The construction process is divided into ontology construction, knowledge fusion, knowledge storage and knowledge reasoning (Liu Zheng, Li Yang, Duan Hong, Liu Yao, & Qin Zhiguang, 2016). The outline of the gene map construction is shown in Figure 1 below:

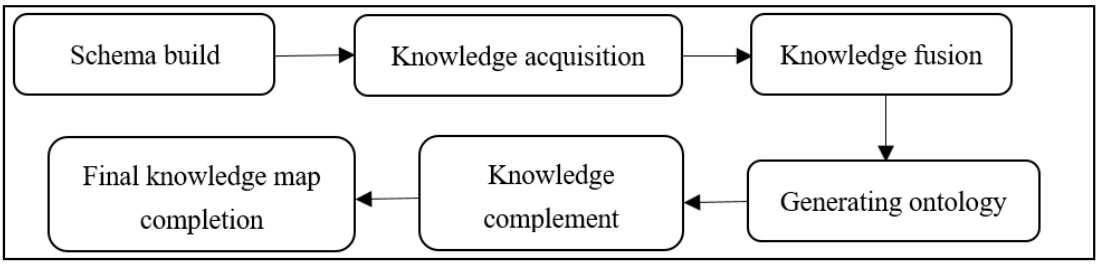

Figure 1 Construction process of genetic map

3.2 Knowledge acquisition

Data Source: The data sources in this article are structured data. The data was mainly from GO (Gene Ontology), OMIM, 'Orphanet, and HPO (Human Phenotype Ontology).

Gene Ontology (Harris, 2004): Gene Ontology (GO) is an ontology widely used in the field of bioinformatics. It mainly includes three branches: biological processes, molecular functions and cell components. The gene ontology is an ontology with a directed acyclic graph.

HPO (Robinson et al., 2008): Human Phenotype Ontology (HPO) aims to provide a standard vocabulary for describing phenotypic abnormalities in human diseases. Each HPO term describes a phenotypic abnormality. The HPO is currently being developed using information obtained from databases such as the medical literature, OMIM, etc. It currently contains about 11,000 entities and more than 115,000 annotations on hereditary diseases.

Orphanet is a knowledge base for the care and treatment of patients with rare diseases by collecting and improving rare diseases to improve diagnostic knowledge. Orphanet aims to provide high quality information on rare diseases. Among them, orphadata has a data set of genes related to rare diseases.

3.3 Build a map

The ontology construction process is generally divided into three types: a top-down approach, a bottom-up approach and a combination of two approaches. This article directly merges the categories, attribute entities and relationships found in the extracted data into the knowledge map. The HPO ontology is used as a third database and Gene Ontology, OMIM and the like are used as structured data to be integrated. Conduct knowledge fusion.

Definition 1: The disease gene entity E refers to various uniquely identifiable pathogenic genes recorded in the knowledge map.

Definition 2: disease gene factual relationship R represents the factual link between different causative genes.

Based on the extraction of relationships from data sources, the following factual relationships of disease-causing genes are finally sorted out.

(1) Has a relationship: indicates the affiliation between entity A and entity B;

(2) Part of relationship: indicates that entity A and entity B are integral and part of the relationship;

(3) Negatively regulates the relationship: indicates that entity A has the function of reverse adjustment of entity B;

(4) Positively regulates: Indicates that entity A has positive adjustment to entity B.

Definition 3: The knowledge map of the disease gene is a directed acyclic graph, stored as a triplet (h,r,t).

3.4 Knowledge Completion

Knowledge representation learning is faced with the representation learning of relations and entities in the knowledge base. It expresses the semantic information representation of entities or relationships by expressing the entities or relationships as low-dimensional real-valued vectors to calculate the semantic associations between entities and relationships. Knowledge representation learning is extremely important for knowledge map completion. Early knowledge representations include

Structured Embedding (SE) (Bordes, Weston, Collobert, & Bengio, 2011) and Single Layer Model (SLM) (Socher, Chen, Manning, & Ng, 2013), Semantic Matching Energy (SME) (Bordes, Glorot, & Weston, 2011). In 2013, Bordes et al. proposed the TransE model (also known as the translation model). This model is based on the phenomenon that Mikolov et al. found that there are lexical semantic relations and translation invariants of syntactic relations. The relationship in the knowledge base is regarded as an entity the translational vector (Mikolov, Chen, Corrado, & Dean, 2013).

In order to solve the inefficiency of the TransE model in dealing with complex relationships (1-N, N-1, N-N relations), a large number of extended models of TransE have been proposed in the past two years. For example, the TransH model proposed by Z Wang et al. in 2014 points out that an entity has different representations under different relationships; the TransSparse model proposed by Ji G et al. in 2016 (Ji, He, Xu, Liu, & Zhao, 2015). Try to use a sparse matrix instead of a dense matrix in the TransR model (Lin, Liu, Zhu, Zhu, & Zhu, 2015) to solve the heterogeneity of entities and relationships (some of the relationships in the knowledge base may be linked to a large number of entities, some relationships may only be associated with a very small number of relationships and imbalances (in some relationships, the types and numbers of head entities and tail entities vary greatly), TransA models (Xiao, Huang, Hao, & Zhu, 2015) try to solve The loss function in TransE is too simple.

(1) Related models

TransE model: Inspired by the translation invariance phenomenon, the TransE model regards the relationship between entities in the knowledge base as some kind of translation between entities and is represented by vectors. Relation l_r can be seen as a translation from head entity vector l_h to tail vector l_t . For each triple (h,r,t) ,in the knowledge base, TransE hopes to satisfy the following relationship: $l_h + l_r \cong l_t$ (1)

TransH model: Since the TransE model cannot be used to deal with complex relationships, the TransH model tries to express the entity structure in different relationships through different forms. For the same entity, it also plays different roles under different relationships. The model first selects a hyperplane F through the relation vector l_r and its normal normal vector w_r, then projects the head entity vector l_h and the tail entity vector l_t in the direction of the normal vector w_r, finally calculates the loss function. TransH enables different entities to have different representations under different relationships, but since the entity vectors are projected into the semantic space of the relationship, they have the same dimensions.

(2) Link Prediction (Knowledge Reasoning）

This project uses the TransH model. This model predicts the physical link of disease gene knowledge maps, and uses HITS@10(%) (ie, 10% before the correct entity) evaluation indicators to achieve a prediction performance of 86%, while predicting the correct evaluation of the results can also achieve a 79% accuracy rate. TransH can be better adapted to the entity prediction link of the constructed knowledge map, so as to achieve the purpose of complementing the knowledge map.

4. The retrieval system based on disease gene knowledge map

Knowledge maps of disease genes increase connectivity in the field of pathogenic gene knowledge, enabling users to browse and query disease-causing gene knowledge resources at the conceptual level. Neoj4 and other map databases can be used to visualize the knowledge base and visualize the way of the semantic web to display more vividly the relationships between entities. At the same time, the search for the predicted pathogenic genes can provide more reliable data for biologists and other users, and it can also provide a new idea for biologists' research on pathogenic genes.

5.Conclusion

This project is based on systematic analysis and in-depth analysis of disease genes and knowledge map construction related literature, making full use of multiple knowledge bases (or ontology) and other data resources in the field of biological genes, using artificial intelligence to construct disease genes. The knowledge map uses the TransH model for knowledge reasoning (Map Completion). Based on the knowledge map and Neoj4 build disease gene knowledge retrieval system, it provides new materials for the related work of pathogenic gene prediction.

References

[1] Bollacker, K., Evans, C., Paritosh, P., Sturge, T., & Taylor, J. (2008). Freebase:a collaboratively created graph database for structuring human knowledge. Paper presented at the SIGMOD Conference.

[2] Bordes, A., Glorot, X., & Weston, J. (2011). Joint Learning of Words and Meaning Representations for Open-Text Semantic Parsing. Paper presented at the International Conference on Artificial Intelligence & Statistics.

[3] Bordes, A., Usunier, N., Garcia-Duran, A., Weston, J., & Yakhnenko, O. (2013). Translating Embeddings for Modeling Multi-relational Data. Paper presented at the International Conference on Neural Information Processing Systems.

[4] Bordes, A., Weston, J., Collobert, R., & Bengio, Y. (2011). Learning Structured Embeddings of Knowledge Bases. Paper presented at the AAAI Conference on Artificial Intelligence, AAAI 2011, San Francisco, California, Usa, August.

[5] Harris, M. A. (2004). The Gene Ontology (GO) database and informatics resource: WCB/McGraw-Hill.

[6] Heilig, R., Eckenberg, R., Petit, J. L., Fonknechten, N., Silva, C. D., Cattolico, L., . . . Ureta-Vidal, A. (2004). International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome. Nature 431: 931-945. Nature, 431(7011), 931-945.

[7] Ji, G., He, S., Xu, L., Liu, K., & Zhao, J. (2015). Knowledge Graph Embedding via Dynamic Mapping Matrix. Paper presented at the Meeting of the Association for Computational Linguistics and the International Joint Conference on Natural Language Processing.

[8] Lin, Y., Liu, Z., Zhu, X., Zhu, X., & Zhu, X. (2015). Learning entity and relation embeddings for knowledge graph completion. Paper presented at the Twenty-Ninth AAAI Conference on Artificial Intelligence.

[9] Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. Computer Science.

[10] Robinson, P. N., Köhler, S., Bauer, S., Seelow, D., Horn, D., & Mundlos, S. (2008). The Human Phenotype Ontology: A Tool for Annotating and Analyzing Human Hereditary Disease. American Journal of Human Genetics, 83(5), 610-615.

[11] Socher, R., Chen, D., Manning, C. D., & Ng, A. Y. (2013). Reasoning with neural tensor networks for knowledge base completion. Paper presented at the International Conference on Neural Information Processing Systems.

[12] Suchanek, F. M., Kasneci, G., & Weikum, G. (2007). Yago:a core of semantic knowledge. Paper presented at the International Conference on World Wide Web.

[13] Walter, S., Unger, C., & Cimiano, P. (2015). DBlexipedia: A Nucleus for a Multilingual Lexical Semantic Web.

[14] Wang, Z., Zhang, J., Feng, J., & Chen, Z. (2014). Knowledge Graph Embedding by Translating on Hyperplanes. Hao AAAI - Association for the Advancement of Artificial Intelligence.

[15] Xiao, H., Huang, M., , Y., & Zhu, X. (2015). TransA: An Adaptive Approach for Knowledge Graph Embedding. Computer Science.

[16] Jiang Xin. (2005). Bioinformatics Database and Its Utilization Methods. Modern Intelligence, 25(6), 185-187.

[17] Liu Yi, Li Yang, Duan Hong, Liu Yao, & Qin Zhiguang. (2016). Review of knowledge map construction techniques. Computer Research and Development, 53(3), 582-600.
[18] Liu Zhiyuan, Sun Maosong, Lin Yankai, & Xie Ruobing. (2016). Progress in knowledge representation learning. Computer Research and Development, 53(2), 247-261.
[19] Wang Zhihong. (2006). An Overview of Knowledge Discovery in the Database. Modernization of the Market (24), 171-172.
[20] Wu Yunbing, Yin Aiying, Lin Kaibiao, Yu Xiaoyan, & Lai Guohua. (2017). Research on the construction of knowledge map based on multiple data sources. Journal of Fuzhou University: Natural Science Edition, 45(3), 329-335.