# SciGraph: A Knowledge Graph Constructed by Function and Topic Annotation of Scientific Papers

#### INTRODUCTION

**Background:** The knowledge of a domain keeps changing under the rapid development of science and technology, researchers have to frequently face new topics and numerous papers while establishing their own studies. On one hand, different types of papers are needed in different stages of a research. On the other hand, the relations among related topics helps researchers building a complete understanding on unfamiliar topics.



- ◆Target: (1) Annotate scientific papers in two dimension, i.e. function and topic, (2) Construct a knowledge graph named SciGraph to organize the annotation results.
- ◆Significance: (1) Use SciGraph to form a literature retrieval system; (2) Help researchers obtain both purpose-oriented papers and the knowledge structure of a domain.

- domain coverage.

$$\begin{split} itf_{i} &= \frac{1}{\left|\sum_{j=1}^{m} \sum_{t \in d_{j}} \{t:t=t_{i}\}\right|}, i \in \{1,2,3,\dots,n\} \\ df_{i} &= \left|\sum_{j=1}^{m} \{j:t_{i} \in d_{j}\}\right|, i \in \{1,2,3,\dots,n\} \\ df &- itf_{i} = df_{i} \times itf_{i}, i \in \{1,2,3,\dots,n\} \end{split} \qquad \mathcal{C}(k_{1},k_{2}) = \begin{cases} \frac{F(k_{1},k_{2})}{F(k_{2})} &, F(k_{1},k_{2}) \geq \varepsilon, F(k_{1}) > F(k_{2}) \\ 0 &, F(k_{1},k_{2}) < \varepsilon \\ -\frac{F(k_{1},k_{2})}{F(k_{1})} &, F(k_{1},k_{2}) \geq \varepsilon, F(k_{1}) \leq F(k_{2}) \end{cases} \end{split}$$

- relations based on the weights above.
- **♦ Knowledge Graph Construction:** This study construct a knowledge graph, SciGraph, to organize the annotated functions and topics of papers. There are four kinds of nodes (total ca. 1.9 million) and four kinds of relations (total ca. 16.4 million) among above nodes.

## Yuchen Yan, Chong Chen

Beijing Normal University, Beijing, China

#### **MATERIALS & METHODS**

◆ **Dataset:** The dataset is composed of ca. 0.9 million Chinese scientific papers with titles, abstracts and keywords, and sampled from 1,203 domains in 11 scientific disciplines in order to keep a reasonable

**Function Annotation:** (1) Propose a DF-ITF method to select the function words with top 1000 DF-ITF values; (2) Cluster the identified function words into six function types: Review & Progress, Demonstration & Comparison, Argumentation & Discussion, Theory & Computation, Technology & Method, and Design & Application; (3) Apply a BERT-based text classification method to annotate a suitable function on scientific papers, the overall accuracy of the model achieves 0.72.

**Topic Annotation:** (1) Using a BERT-based token classification to extract topic keywords from abstracts of scientific papers, the F1 value reaches 0.475; (2) Apply a term-occurrence-based method to calculate the hyponym relation weight  $C(k_1, k_2)$  of topic keyword pairs  $k_1$  and  $k_2$ ; (3) Identify and rank the hyponym



This system contains a distinguishing function, i.e. displaying knowledge graph, which meet the need of explorative retrieval particularly.



### **APPLICATION DEMO**

In order to show the significance of the proposed solution on explorative retrieval, this study design and construct a scientific paper retrieval system.

	apers Retrieval	System				Function Annotation 设计与应用
Scientific Papers Retrie	eval					v
Title						Topic Annotation 卷积神经网络,深度学习,调制样式
Abstract	自动调制样式识别分类是解调前的重 映射成星座图的具有不同调制样式的 ≥5dB时,识别率可达97.99%,信噪比	重要步骤,在频谱管理、 ウ通信信号馈送进神經 ;≥9dB时,识别率可达	、认知无线电、智能调制解调器 圣网络,从而达到通信信号调试档 ;100%.	8、监视和干扰识别等许多应用中发 并式识别分类的目的.基于实验目的;	就择着重要作用.深度学习具有强大的分类能力 提出一种改进的卷积神经网络结构可实现对七	,基于深度学习中的卷积神经网络,将 ;种不同的调制样式的分类,在信噪比
Topic	深度学习					
CLC						
Journal						
Function	Z Review & Progress Demo	onstration & Compa	arison Argumentation	& Discussion 🗌 Theory &	Computation	Design & Application
Function Hyponym	Review & Progress Demo 算度学习技术 算度学习模型 算度	onstration & Compa 度学习方法 深度学习 Retrieva	arison Argumentation 网络 深度学习算法 长翅的	& Discussion Theory & 財记忆网络 长知时记忆 网络初 Function Annotation Topic A	Computation Technology & Methon	d 🗌 Design & Application
Function Hyponym Title	Review & Progress Demo 原度学习技术 深度学习機型 深度	onstration & Compa 建学习方法 深度学习 Retrieva Authors	arison Argumentation 网络 深度学习算法 长短 I Knowledge Graph Journal	& Discussion 口 Theory & d 時记忆開始 长知時记忆 開始如 Function Annotation Topic A CLC	Computation Technology & Metho ILS 2574922918 Page Numeration Keyphrase	d Design & Application
Function Hyponym Title > 基于机器学习	Review & Progress Demc 深度学习技术 深度学习概型 深度	atthors 案学习方法 深度学习 Retrieva Authors 表洪,胡昌华, 司小胜	arison Argumentation 网络 深度学习算法 长翅 I Knowledge Graph Journal 《机械工程学报》	& Discussion Theory & d 時记忆開始 任初時记忆 网络如 Function Annotation Topic A CLC V44.就天仪来、航天器设备	Computation Technology & Methon IES 在初神经网络 降雄 Annotation Keyphrase 支持向量机 机器学习 神经网络 深度学习 教会本会预测 教会来会 寿会	d Design & Application Function 5 技术与方法 经送与进展
Function Hyponym Title > 基于机器学习 ) 单幅图像例体	2 Review & Progress     Demo       深度学习技术     深度学习技术     深度学习技术       的设备剩余寿命预测方法结述     目的设备剩余寿命预测方法结述	Retrieva 文法 構築 定 定 定 定 定 定 定 定 定 定 定 定 定	Argumentation 同緒 深度学习算法 长期 I Knowledge Graph Journal 《机械工程学报》	& Discussion Theory & i 對记忆規模 任相對记忆 网络说 Function Annotation Topic A CLC V44.集天仪表、集天卷设备 TP37.多级体技术均多级体;	Computation Technology & Methon IIS 在初神经网络 算槍 Windtation Keyphrase 支持向最低 机器学习 神经网络 深度学习 多指曲度 计算机规定 姿态估计 深度学习	d Design & Application Function 2 技术与方法 或线与波展 2 和体目标 技术与方法 编送与波展
Function Hyponym > 基于机器学习 > 单幅图像刚体 > 基于卷积神经	2 Review & Progress     Demo       漢度学习技术     漢度学习电型     漢集       的设备剩余寿命预测方法综述        国标姿态估计方法综述        网络的光学温感目标检测研究进展	#学习方法 深度学习 使度学习方法 深度学习 Retrieva Authors 製川,胡昰 報子,一般小 平,万字强 影楽観,総云 飞,马中祺	Argumentation 深度学习算法 长知 《 Knowledge Graph Journal 《 机械工程学报》 《 化、磁路表面形学报》	& Discussion □ Theory & H 封记忆興時 任知時记忆 网络说 Function Annotation Topic A CLC V44.惹天仪表、航天電波番 [17937.多媒体技术与多媒体: P40一般理论与方法	Computation     Technology & Methodists       IEE     各駅神经网络     降信       Windtation         Keyphrase         文材向量机     和余寺会     寿命       多自由度     计算机极度     姿态估计     深度学习       光学 運感     光学運感     各駅神经网络     目 深度学习	d Design & Application Function 2 授术与方法 经关号过展 2 期件目标 授术与方法 经过与过展
Function Hyponym Title > 基于机器学习 之 单磁图像别体 > 基于教职神经	2     Review & Progress     Demo       深度学习技术     深度学习模型     深度       2     的设备剩余寿命预测方法综述        2     的设备剩余寿命预测方法综述        2     回路的光学温感目标检测研究进展        2     双其故障诊断应用分析与展望	ま学习方法 深度学习 深度学习 Retrieva Authors 表現、胡島 年、 司小班	Argumentation 同時的 定度学习算法 长期 人 Knowledge Graph Journal 《机械工程学报》 《 《 中国图象图形学报》 《 《 而 安 交 逸 大学学报》	& Discussion □ Theory & i 時記忆発播 长期時记忆 网络第 Function Annotation Topic A CLC · · · · · · · · · · · · · · · · · ·	Computation     Technology & Methon       IES     各科神经网络     開始       Windtation         Kayphraso         支持向量机     机器学习     神经网络     項度学习       教念寿会規測     第余会 考会         支持由量化     計算机规定     要态估计     深度学习       光学     運送     光学運送     各科神经网络     目       現成学会     大致重     政務诊断     須度学习	Design & Application   Function   現代目示   日