Paper Pal: 一个中英文论文及其代码大数据搜索平台

Paper Pal: a big data search platform combining Chinese and English papers and codes

余万（1997- ），男，复旦大学计算机科学技术学院、上海市数据科学重点实验室硕士生，主要研究方向为数据挖掘及其应用。

付聿炜（1994- ），男，复旦大学计算机科学技术学院、上海市数据科学重点实验室硕士生，主要研究方向为异质网络、推荐系统。

熊赟（1980- ），女，复旦大学计算机科学技术学院教授、博士生导师，上海市数据科学重点实验室副主任。从2004年起从事数据领域方面的研究工作，作为项目负责人，主持多项国家自然科学基金项目、上海市科学技术委员会发展基金项目以及企业合作项目。在国际权威期刊和会议论文集上发表论文80余篇，出版著作3本。目前主要研究方向为数据科学和大数据。

朱扬勇（1963- ），男，复旦大学计算机科学技术学院教授、博士生导师，上海市数据科学重点实验室主任。从1989年起从事数据领域的研究工作，1996年开始从事数据挖掘研究工作，2004年开始从事数据科学研究工作，是国内最早一批从事数据挖掘研究工作的学者和国际数据科学研究工作的主要倡导者之一。2009年发表了数据科学论文“Data explosion, data nature and dataology”，并出版第一本数据科学专著《数据学》。主持国家自然科学基金项目、国家863计划项目、上海市科学技术委员会重点项目等数十项研究课题，曾获上海市科技进步奖一、二、三等奖。在国内外权威期刊或会议上发表论文150余篇，出版专著2本，教材3本。目前主要研究方向为数据科学和大数据。

中图分类号：TP311　　文献标识码：A　　doi: 10.11959/j.issn.2096-0271.2020049

1 引言

在开展科研工作的过程中，科研人员需要从大量实时更新的论文中持续地跟踪学术界前沿的发展情况，学习最新研究成果。近年来，人工智能（artificial intelligence，AI）、数据挖掘等领域受到的关注度不断增加，相关会议的论文数量呈爆发式增长。**图1**显示了arXiv[1]数据库中2010—2019年AI领域的论文增长情况[2]。但是，巨大的论文数量导致科研人员搜索论文的过程中耗费了大量的时间。

目前，已经有很多论文搜索引擎，如Microsoft Academic[3]、Arxiv Sanity Preserver、Papers With Code以及AMiner[4]等。其中，Microsoft Academic根据研究领域对论文进行了分类，并提供了论文的全文链接、所发表的会议或期刊、引用的参考文献等；Arxiv Sanity Preserver提供了arXiv上论文的浏览、搜索和排序功能，并根据用户收藏的论文，使用TF-IDF[5]和支持向量机（support vector machine，SVM）[6]实现论文推荐。对于计算机领域的科研人员，论文中提出的算法、模型的代码是相当重要的学习资源，能够让人更直观、快速地理解和掌握一个新算法或新模型[7]，但Microsoft Academic和Arxiv Sanity Preserver等未提供代码信息。在众多代码平台中，GitHub[8]成为目前非常有代表性的代码平台。但是，在搜索论文和对应的代码时，科研人员需要在不同的搜索平台上来回切换以获取论文和代码，这无疑增加了科研工作的时间。

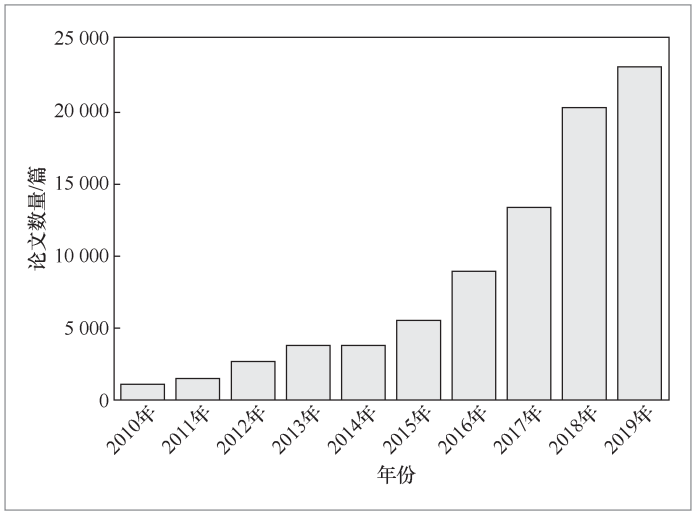


图1 2010—2019年arXiv数据库中AI领域的论文增长情况

为了解决论文和论文代码在空间上的差异问题，Atlas ML推出一个免费、开源的机器学习领域的论文和代码分享平台——Papers With Code，该平台不提供计算机领域的中文论文。AMiner是目前功能较全的研究者、论文搜索平台，其构建的主要目标是通过整合多源数据提供研究者搜索分析功能来构建研究者网络和学术论文网络[9]。该平台也提供中英文论文的搜索功能，并包含部分可人工编辑的论文相应的代码链接。

不同于AMiner平台的构建目标，本文聚焦中国计算机领域的科研人员在搜索论文中的实际需求，以“中国计算机学会（China Computer Federation，CCF）推荐分区论文+代码+中文期刊+推荐”为定位，设计和实现了一个使用友好、免费、开源的计算机领域论文与代码搜索系统——Paper Pal。

Paper Pal针对中国计算机领域的科研人员需求，按照CCF推荐论文分区对平台中的论文进行分类，提供方便的选项卡和搜索支持，平台功能更加聚焦。目前，Paper Pal共收录英文文献29 507篇、中文文献2 130篇以及代码6 147份，覆盖人工智能、数据挖掘领域的CCF分区的35个A类和B类会议以及四大计算机领域中文

期刊(《计算机学报》《软件学报》《计算机研究与发展》《大数据》)。同时,平台具有可扩展性,将持续收集整合新的会议和期刊的论文。

此外,用户也可以采用关键词、期刊名、会议名等方式进行论文搜索。Paper Pal还提供论文收藏功能,以便用户记录、整理以及追溯,同时将用户收藏的论文作为用户的行为数据来源,利用系统内置的论文推荐功能,推测用户可能感兴趣的论文,帮助用户在更短的时间内找到所需要的资料,进一步提升用户的使用体验。**表1**对Paper Pal和上述其他论文搜索平台进行了对比。

2 相关技术

Paper Pal平台负责对论文及其代码进行收集、存储并提供搜索和推送功能。这里需要解决几个问题:第一,数据是多来源的,如何合理地进行数据整合,以提升平台数据质量是基础;第二,论文数据是文本类型,如何有效地进行存储和预处理是核心;第三,平台中既有中文论文又有英文论文,如何提供精准搜索是关键。针对上述问题,本文采用了当前大数据领域中主流的数据获取、数据存储和数据检索技术,并根据本系统的特点进行了改进,包括在MongoDB与Elasticsearch之间进行数据同步,从而实现在利用MongoDB数据存储的优势和Elasticsearch强大的中文搜索能力的同时,保证搜索结果与数据存储更新的同步。此外,采用面向异质网络的推荐模型对用户的搜索行为进行分析,以实现论文推送。

2.1 数据获取与存储

实现Paper Pal的第一步是收集论文及其相关数据,目前本系统收集了CCF推荐分区A类、B类会议近3年发表的人工智能、数据挖掘领域的论文。数据来自DBLP[10]上论文所在的期刊、会议的详细信息。同时,从Microsoft Academic上获取了论文被引用的次数。综合上述信息,系统提供的论文相关信息包括论文的标题、作者、出版日期、论文PDF文档链接和被引

表1 Paper Pal 与各论文搜索平台的对比

对比项	Paper Pal	Papers With Code	AMiner	Microsoft Academic	Arxiv Sanity Preserver
构建目标	面向中国计算机领域的科研人员,聚焦CCF推荐分区会议及中文期刊论文和代码	主要提供来自arXiv的英文论文及其代码	收集中英文论文,构建研究者网络和学术图谱	英文论文搜索	收集物理学、数学、计算机科学与生物学论文预印本
搜索功能	会议、期刊选项卡、关键词、条件查询	关键词、条件查询	关键词、条件查询	关键词、条件查询	关键词、条件查询
是否按CCF推荐分区	是	否	否	否	否
是否含中文期刊论文	是	否	是	是(时效性不强)	否
是否提供论文引用	是	否	是	是	否
是否附带论文代码	是	是	是	否	否
是否有推荐功能	是	否	是	否	否
是否提供公开数据集	是	是	是	否	否

用的次数等。除英文论文外，Paper Pal还从计算机领域的中文期刊中获取了中文论文数据。

获取到论文信息后，进一步整合论文中介绍的模型和算法的相关代码。系统将GitHub当作Paper Pal的代码数据来源，通过GitHub提供的API来获取代码数据。虽然有些论文没有论文原作者公布的代码，但是会有其他研究人员在GitHub上分享实现的代码。

收集完论文数据和对应的代码之后，将其存储到数据库中。本系统使用MongoDB提供数据存储和管理服务。每篇英文论文的记录有11个属性，分别为：论文的ID、标题、摘要、作者、发布日期、代码链接、PDF链接、关键词、被引用次数、发表会议或期刊、发表年份。

2.2 中英文论文搜索方法

Paper Pal收集的论文包括中文论文和英文论文。为实现更高效、准确、方便的中英文检索功能，Paper Pal选取Elasticsearch[11]作为搜索引擎。Elasticsearch是一个开源的、基于Lucene的分布式数据搜索引擎，能够提供快速的检索功能，具有易扩展、近实时的特点。Elasticsearch的倒排索引功能能够有效地提高多条件查询的检索效率；Elasticsearch支持中文分词插件IK Analyzer，能够更好、更方便地满足Paper Pal对中文文献的检索需求。除此之外，Elasticsearch还有与之配套的可视化工具Kibana和日志收集分析工具Logstash，能够为Paper Pal提供日志收集、文本检索和数据可视化分析整套流程的服务[12]。

但Elasticsearch容易因为软硬件崩溃而造成数据丢失且无法恢复，因此Elasticsearch通常与关系型数据库或非关系型数据库配合使用，其中数据库作为持久化存储组件提供约束限制和系统鲁棒性保证，而Elasticsearch基于数据内容实现复杂的搜索查询。Paper Pal的数据被存储在MongoDB中，在本系统中，笔者把MongoDB的论文数据同步到Elasticsearch中，并实时监听MongoDB中数据的更新情况。如**图2**所示，Paper Pal使用Mongo-connector来跟踪事先建立好的MongoDB Replica Set的oplog（operations log），利用Mongo-connector的文档管理器Elastic2-doc-manager将MongoDB的数据导入Elasticserach，并实时监听oplog的变化，以保持Elasticsearch与MongoDB之间数据的同步。

2.3 论文推荐方法

考虑到目前收集的用户数据有限，目前Paper Pal使用与Arxiv Sanity Preserver相似的基于内容的推荐方法，即根据用户收藏的论文的标题与摘要，使用TF-IDF和SVM将论文的词频等作为特征来计算其他论文和用户收藏的论文在词的语义上的相似度。同时，Paper Pal系统内置了笔者提出的基于异质网络表示学习的基于元路径增强的图注意力编码（metapath enhanced graph attention encoder，MEGAE）[13]模型，模型框架如**图3**所示。该模型将论文、用户看成一个异质网络，将用户搜索以及收藏的论文作为用户和论文之间的边，当用户注册并登录Paper Pal后，Paper Pal会将用户收藏和浏览的论文信息记入数据库，这些数据将被用来更新网络，为推荐功能积累数据来源。例如，当用户A看了论文B之后，Paper Pal会在异质网络中为用户A和论文B添加一条连边。Paper Pal使用MEGAE模型学习异质网络中每个不同节点的低维向量表示[14]和异质

网络结构信息[15]，捕捉用户的兴趣点，从而为用户推荐论文。比起单纯使用词频作为特征进行推荐，MEGAE模型不仅能捕捉到异质网络的结构信息，同时还能学习到异质网络中隐含的语义关系，实现更精准的个性化推荐。根据本系统的特点，即论文具有CCF分区信息，发表论文的会议或期刊所属的CCF分区和论文领域可以作为论文的标签加入论文节点的属性中，即将MEGAE模型应用到考虑节点属性的属性网络图中。

3 平台效果

Paper Pal平台为中国计算机领域的科研工作者提供了“分区搜索”功能，即直接进入CCF推荐分区会议或中文期刊进行搜索（如**图4**所示）。用户可以选择浏览CCF推荐分区会议或中文期刊的论文，系统根据用户的选择显示相应的论文列表。论文列表包括论文的标题、作者、发表日期、摘要、PDF文档链接、代码链接以及被引次数等信息。考虑到存在具体某一期刊/会议论文数量多并且用户只想搜索该期刊/会议下的论文的情况，Paper Pal为用户提供两种搜索范围，一种是在所有期刊/会议下进行搜索，另一种是在某个特定的期刊/会议下进行搜索。“分区搜索”是区分于其他平台的重要功能。因为用户对高质量论文的关注度通常更高，所以本功能通过给出中国计算机学会的高质量会议推荐列表及其中的论文，为用户提供直接的搜索服务。而在现有其他平台上，用户必须先去查阅哪些会议在中国计算机学会的推荐列表中，然后再到搜索平台中用关键词进行检索。因此，本平台将大幅减少用户在搜索高质量论文（计算机学会推荐列表中的会议论文）时耗费的时间。此外，用户耗费相当时间查阅到所需的会议名之后，在现有其他平台上将会议名作为关键词进行搜索时，搜索结果会出现

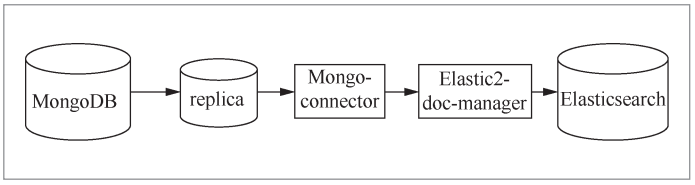


图 2 将 MongoDB 的数据同步到 Elasticsearch 中

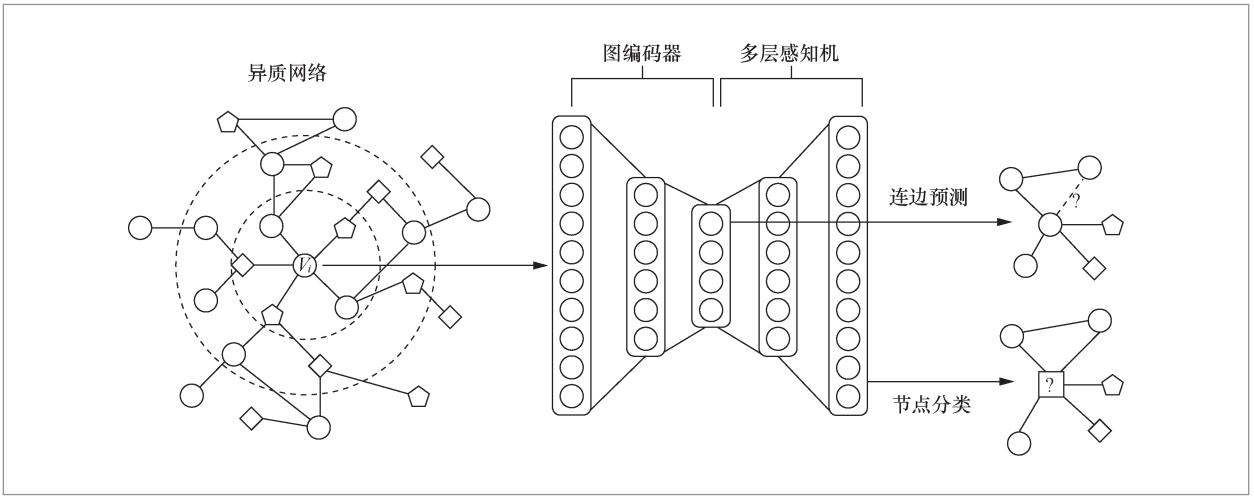


图 3 MEGAE 模型框架

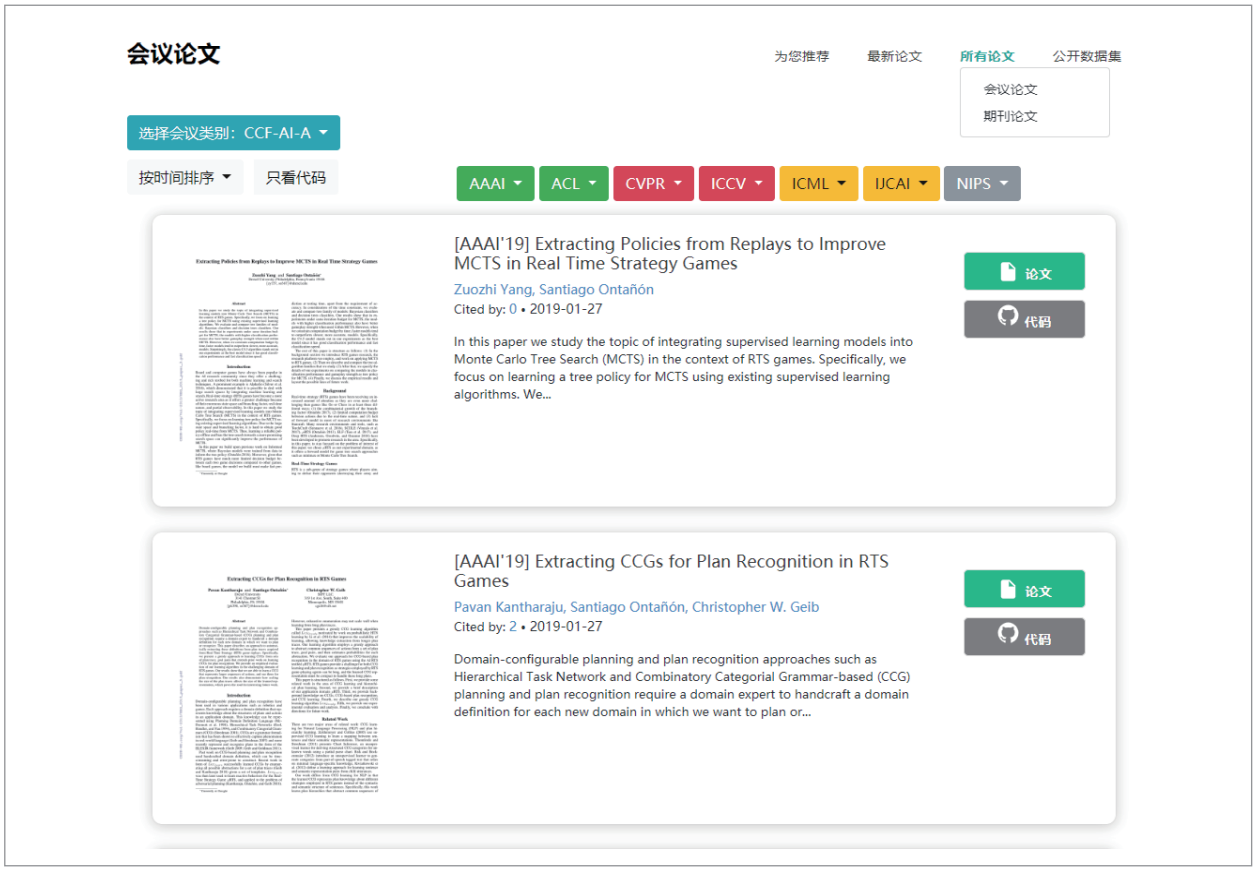
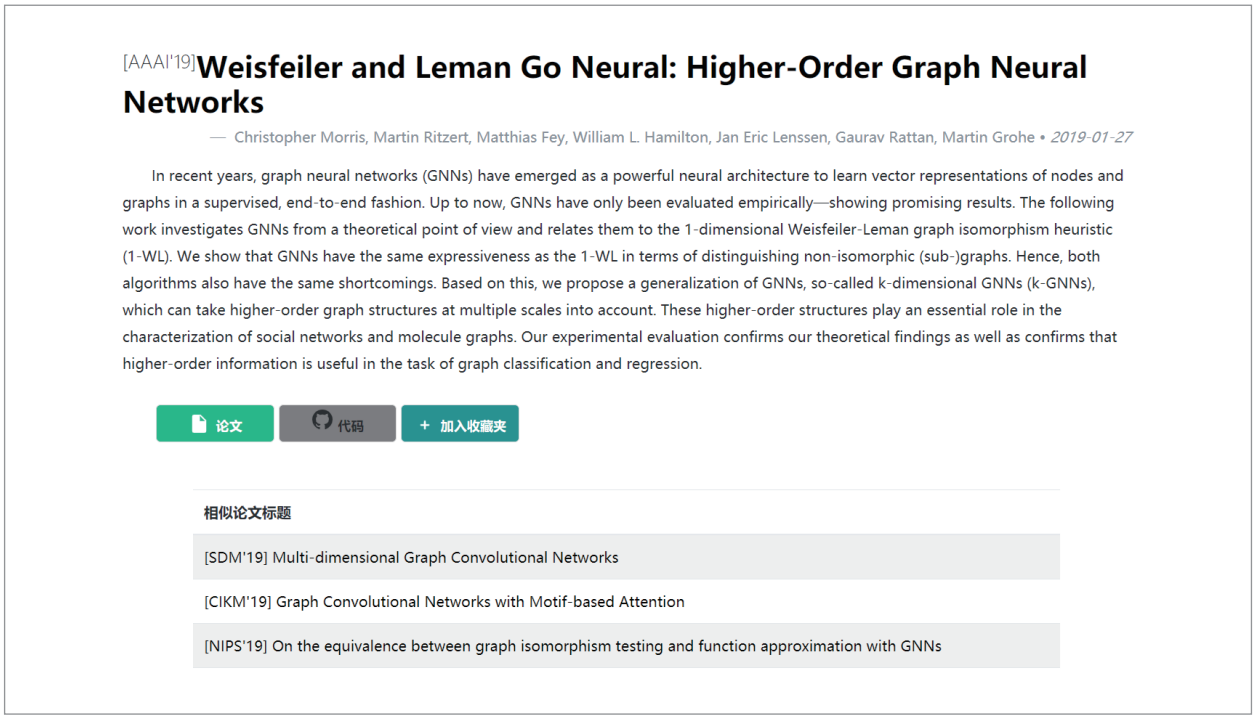


图 4 Paper Pal 的“分区搜索”页面

偏差。例如international conference on machine learning（ICML）中有“machine learning”，若将“machine learning”作为关键词进行搜索，将会把该词作为标题或摘要等中的匹配词返回，而不是搜索ICML。最后，如果在现有其他平台上直接使用会议名的缩写来搜索会议，对搜索质量将是更大的挑战。因此，本平台的搜索聚焦关键词与论文主题等的匹配度，而不需要考虑以会议名为关键词的匹配，所以，本平台具有更高的精准度。

当用户查阅到自己感兴趣的论文时，可以进入论文的详情页面，将论文添加到收藏夹中。Paper Pal根据论文的标题和摘要使用TF-IDF和SVM生成该论文的相似论文目录。用户可在论文的详情页面（如图5所示），进一步查看与当前论文相似的论文。Paper Pal基于MEGAE模型的论文推荐功能需用户注册、登录，并且在平台积累到一定数量的用户收藏数据后才能使用。MEGAE模型使用图注意力编码器来捕捉网络结构的信息，能够增强模型的可解释性，同时还能学习到由论文、作者、会议/期刊等构成的异质网络中隐含的语义关系，实现更精准的个性化推荐。例如，可以根据论文是否具有合作者或论文是否发表在不同会议上等不同的条件，实现不同语义路径下的推荐。其生成的推荐目录可在“为您推荐”版块中查阅。

Paper Pal也将持续收集和更新一系列公开的数据集，并根据不同的研究方向对数据集进行划分（如图6所示）。



[AAAI'19]**Weisfeiler and Leman Go Neural: Higher-Order Graph Neural Networks**

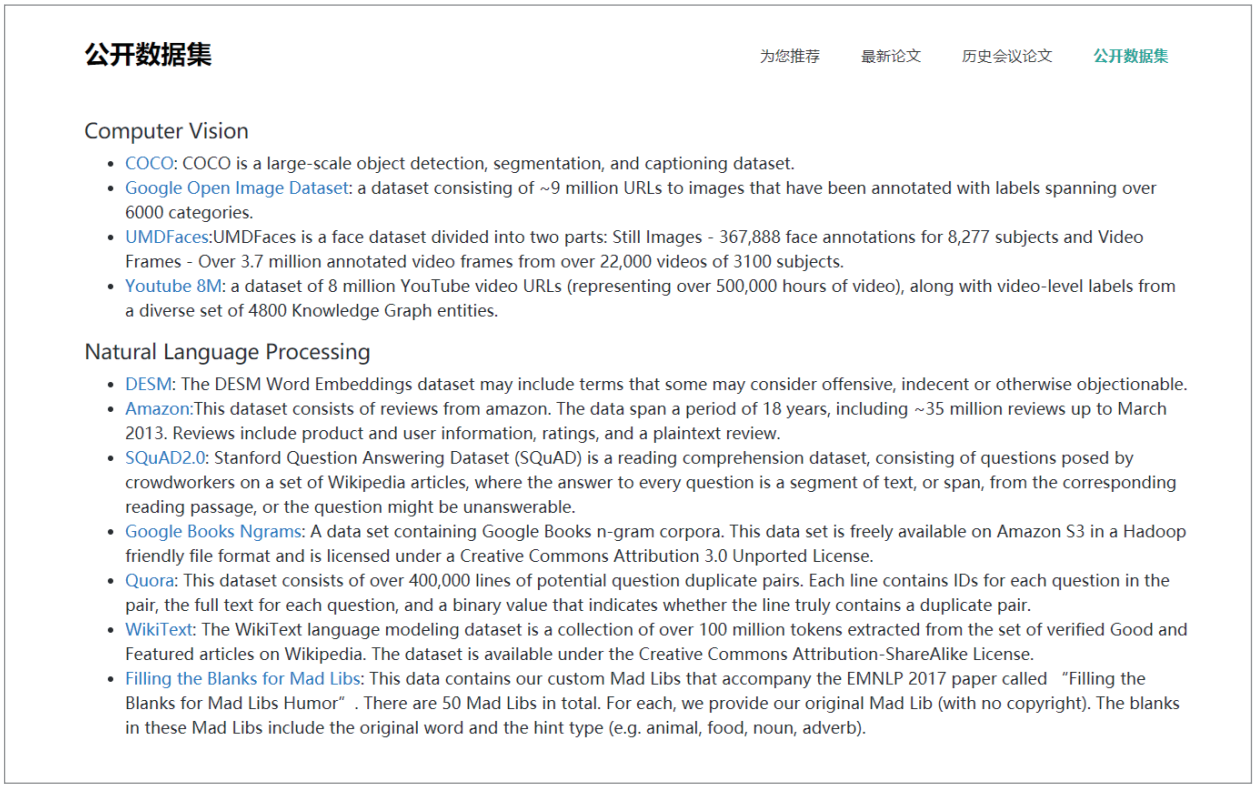
— Christopher Morris, Martin Ritzert, Matthias Fey, William L. Hamilton, Jan Eric Lenssen, Gaurav Rattan, Martin Grohe • *2019-01-27*

In recent years, graph neural networks (GNNs) have emerged as a powerful neural architecture to learn vector representations of nodes and graphs in a supervised, end-to-end fashion. Up to now, GNNs have only been evaluated empirically—showing promising results. The following work investigates GNNs from a theoretical point of view and relates them to the 1-dimensional Weisfeiler-Leman graph isomorphism heuristic (1-WL). We show that GNNs have the same expressiveness as the 1-WL in terms of distinguishing non-isomorphic (sub-)graphs. Hence, both algorithms also have the same shortcomings. Based on this, we propose a generalization of GNNs, so-called k-dimensional GNNs (k-GNNs), which can take higher-order graph structures at multiple scales into account. These higher-order structures play an essential role in the characterization of social networks and molecule graphs. Our experimental evaluation confirms our theoretical findings as well as confirms that higher-order information is useful in the task of graph classification and regression.

论文 代码 + 加入收藏夹

相似论文标题
[SDM'19] Multi-dimensional Graph Convolutional Networks
[CIKM'19] Graph Convolutional Networks with Motif-based Attention
[NIPS'19] On the equivalence between graph isomorphism testing and function approximation with GNNs

图 5　论文的详情页面



公开数据集

为您推荐　最新论文　历史会议论文　公开数据集

Computer Vision

- COCO: COCO is a large-scale object detection, segmentation, and captioning dataset.
- Google Open Image Dataset: a dataset consisting of ~9 million URLs to images that have been annotated with labels spanning over 6000 categories.
- UMDFaces:UMDFaces is a face dataset divided into two parts: Still Images - 367,888 face annotations for 8,277 subjects and Video Frames - Over 3.7 million annotated video frames from over 22,000 videos of 3100 subjects.
- Youtube 8M: a dataset of 8 million YouTube video URLs (representing over 500,000 hours of video), along with video-level labels from a diverse set of 4800 Knowledge Graph entities.

Natural Language Processing

- DESM: The DESM Word Embeddings dataset may include terms that some may consider offensive, indecent or otherwise objectionable.
- Amazon:This dataset consists of reviews from amazon. The data span a period of 18 years, including ~35 million reviews up to March 2013. Reviews include product and user information, ratings, and a plaintext review.
- SQuAD2.0: Stanford Question Answering Dataset (SQuAD) is a reading comprehension dataset, consisting of questions posed by crowdworkers on a set of Wikipedia articles, where the answer to every question is a segment of text, or span, from the corresponding reading passage, or the question might be unanswerable.
- Google Books Ngrams: A data set containing Google Books n-gram corpora. This data set is freely available on Amazon S3 in a Hadoop friendly file format and is licensed under a Creative Commons Attribution 3.0 Unported License.
- Quora: This dataset consists of over 400,000 lines of potential question duplicate pairs. Each line contains IDs for each question in the pair, the full text for each question, and a binary value that indicates whether the line truly contains a duplicate pair.
- WikiText: The WikiText language modeling dataset is a collection of over 100 million tokens extracted from the set of verified Good and Featured articles on Wikipedia. The dataset is available under the Creative Commons Attribution-ShareAlike License.
- Filling the Blanks for Mad Libs: This data contains our custom Mad Libs that accompany the EMNLP 2017 paper called "Filling the Blanks for Mad Libs Humor". There are 50 Mad Libs in total. For each, we provide our original Mad Lib (with no copyright). The blanks in these Mad Libs include the original word and the hint type (e.g. animal, food, noun, adverb).

图 6　公开数据集页面

4 结束语

本文设计和实现了一个论文及其代码大数据搜索系统——Paper Pal，旨在为中国计算机领域的科研人员提供一个功能更全面的中英文论文及其代码大数据搜索工具。该平台基于多源数据获取、MongoDB数据库存储、非结构化文本抽取转换和Elasticsearch中文数据检索等方法和技术，整合了CCF推荐分区会议和部分国内计算机领域的中文期刊的论文及其已公开在GitHub上的代码，并提供论文及其代码大数据搜索功能；还采用面向异质网络的推荐模型实现用户搜索行为分析，为用户推送感兴趣的论文。Paper Pal平台将大幅缩短科研人员查找文献的时间，帮助科研人员在更短的时间内更有效地获取更多、更全面的资料，并且该平台中积累的计算机领域高质量中英文论文、代码及其数据集形成了科研成果研究的大数据资源库，为科研大数据研究提供了丰富的数据基础，也为科研趋势分析研究提供了数据支持，对持续开展科研领域的成果进展研究具有重要意义。

参考文献:

[1] GINSPARG P. ArXiv at 20[J]. Nature, 2011, 476(7359): 145-147.

[2] PERRAULT R, SHOHAM Y, BRYNJOLFSSON E, et al. The AI Index 2019 Annual Report[R]. 2019.

[3] SINHA A, SHEN Z, SONG Y, et al. An overview of Microsoft Academic Service (MAS) and applications[C]// The 24th International Conference on World Wide Web. [S.l.:s.n.], 2015: 243-246.

[4] WAN H Y, ZHANG Y T, ZHNAG J, et al. AMiner: search and mining of academic social networks[J]. Data Intelligence, 2019, 1(1): 58-76.

[5] RAMOS J. Using TF-IDF to determine word relevance in document queries[C]// The 1st Instructional Conference on Machine Learning. [S.l.:s.n.], 2003: 133-142.

[6] CHANG C C, LIN C J. LIBSVM: a library for support vector machines[J]. ACM Transactions on Intelligent Systems and Technology, 2011, 2(3): 1-27.

[7] VAUGHAN R, WAWERLA J. Publishing identifiable experiment code and configuration is important, good and easy[R]. 2012.

[8] BLISCHAK J D, DAVENPORT E R, WILSON G. A quick introduction to version control with Git and GitHub[J]. PLoS Computational Biology, 2016, 12(1): e1004668.

[9] TANG J. AMiner: toward understanding big scholar data[C]// The 9th ACM International Conference on Web Search and Data Mining. New York: ACM Press, 2016: 467.

[10] LEY M. DBLP: some lessons learned[J]. Proceedings of the VLDB Endowment, 2009, 2(2): 1493-1500.

[11] GORMLEY C, TONG Z. Elasticsearch: the definitive guide: a distributed real-time search and analytics engine[M]. New York: O' Reilly Media, Inc., 2015.

[12] LAHMADI A, BECK F. Powering monitoring analytics with ELK stack[C]// International Conference on Autonomous Infrastructure, Management and Security (AIMS). [S.l.:s.n], 2015.

[13] FU Y W, XIONG Y, YU P S, et al. Metapath enhanced graph attention encoder for HINs representation learning[C]// 2019 IEEE International Conference on Big Data (Big Data). Piscataway: IEEE Press, 2019: 1103-1110.

[14] BORDES A, USUNIER N, GARCIA-DURAN A, et al. Translating embeddings for modeling multi-relational data[C]// Advances in Neural Information Processing Systems. [S.l.:s.n.], 2013: 2787-2795.

[15] VELIČKOVIĆ P, CUCURULL G, CASANOVA A, et al. Graph attention networks[J]. arXiv preprint, 2017, arXiv: 1710.10903. □

大数据

BIG DATA RESEARCH

ISSN 2096-0271

邮发代号：2-537　国外代号：C9118　定价：35.00元