# 万物互联：学术数据的互联、挖掘与可视化

张秋颖，周乐，唐静瑶，傅洛伊，王新兵
（上海交通大学，上海 200240）

**摘　要**：随着物联网的不断发展，“物”的概念已扩展至学术数据领域。由于物联网节点的海量性以及节点关系的复杂性，用户很难直接从互联的学术数据中获得对所需信息的进一步分析。AceMap 作为一个学术搜索系统，为了能够帮助用户获得全方位的学术信息，通过自主研发的 AceKG 学术知识图谱，为用户提供了个性化查询以及实时生成结果的服务；同时，以学术地图（如论文地图、学者地图等）的方式直观呈现学术数据之间的关系，帮助用户高效获取所需信息。

**关键词**：物联网；学术大数据；可视化；知识图谱
**中图分类号**：TP391
**文献标识码**：A
**doi:** 10.11959/j.issn.2096-3750.2018.00074

# Internet of everything: interconnection, mining and visualization of academic data

ZHANG Qiuying, ZHOU Le, TANG Jingyao, FU Luoyi, WANG Xinbing
Shanghai Jiao Tong University, Shanghai 200240, China

**Abstract:** With the continuous development of the Internet of things, the concept of “things” has also expanded to academic data. Due to the massiveness of IoT node and the complexity of node relationships, it is difficult for users to obtain further analysis of the required information directly from these interconnected academic data. As an academic search system, in order to help users obtain comprehensive academic information, personalized inquiry and real-time results generation services were provided to users by AceMap through the self-developed AceKG academic knowledge map. At the same time, AceMap presents the relationship between academic data visually in the form of academic maps (such as paper maps, author maps, etc.), and users are helped to get the information they need efficiently.

**Key words:** Internet of things, academic big data, visualization, knowledge map

## 1　引言

随着物联网的不断发展，“物”的概念已扩展至学术数据领域。物联网的主要特征之一是节点的海量性，在学术数据方面，除了论文、作者之外，会议、机构、研究领域以及期刊等也是组成学术物联网的数据节点，这些节点不仅数量规模庞大，关系复杂度也远超过简单的“论文—作者”从属关系。然而，原始的学术数据中，每篇论文、每位学者、每个会议等都是独立的节点，它们之间的关系都不够明朗。此外，尽管现有的学术搜索引擎已经能精确地给出查询结果，但是缺乏对用户搜索结果的进一步分析，用户难以获得除了论文本身之外的信息，如不同论文、不同作者之间的关系。可见，为了能够从这些数据中获得更多可利用信息，还需对其相互关系进行进一步的挖掘分析，并以直观的方式呈现。基于此物联网发展需求，以自主研发的 AceKG 学术知识图谱为数据结构，以多样化的学术图谱为表现方式，开发了 AceMap，即一个面向学术大数据的可视化学术搜索系统。

AceMap 侧重于学术数据的挖掘和可视化。在数据挖掘方面，AceMap 构建了 AceKG 学术知识图

**收稿日期**：2018-08-09；**修回日期**：2018-10-29

谱，由于其完全以三元组的形式组织，具有多个实体类别和关系类型，存储的数据庞大且全面，因此便于数据挖掘和处理。在可视化方面，AceMap 以展现学术地图（包括论文地图、合作关系地图等）为主，辅以作者研究兴趣、论文引用等其他量化图表，为用户提供全方位的学术信息。

## 2　知识图谱

知识图谱最早由 Google 在 2012 年提出[2]，是基于图的语义网络，通过一个三元组（实体—关系—实体或实体—属性—值），把实体关系连接成一个网络，可以让计算机更好地存储和管理各种信息。目前，一些常见的主流知识图谱包括 YAGO[3]、NELL[4]、DBpedia[5]、DeepDive[6]等。

在为物联网系统建立数据库时，开发者一般考虑使用 MySQL、Oracle 等常见的关系型数据库，AceMap 在构建之初采用 MySQL 数据库存储数据。然而使用这种数据库需要多表联合查询，查询量较大、花费时间长、可扩展性受限，因此大部分学术地图需要提前生成，难以根据用户的不同需求生成个性化的地图。于是，在此基础上，AceMap 开发了基于知识图谱的图数据库即 AceKG，进行信息存储。

在 AceKG 中，所有对象都表示为一个实体，AceMap 定义了 5 类学术实体：论文、作者、科研机构、研究领域和发表场所（期刊或会议）。对于每个实体，通常有数字、日期、字符串等属性。AceKG 主体结构如图 1 所示，AceKG 存储的知识被描述为三元组，三元组由主语、谓语和宾语构成。

由物联网的任一领域抽象出的各种实体，不可能提前知道所有实体之间的关系。对于 AceKG 来讲，通过简单分类获得的现有已知关系是远远不够的，因此需要进一步挖掘其相互关系。基于一定规则的知识图谱推理，是一种典型、关键的方法。AceMap 设计的一种 AceKG 的推理规则如图 2 所示，图 2 中实线表示已有的关系，虚线表示可以推理出的新关系。通过这些规则，给定三元组中任意两个元素可以预测另一元素的情况，最终在 AceKG 中不断构造新的三元组，达到数据挖掘的目的。

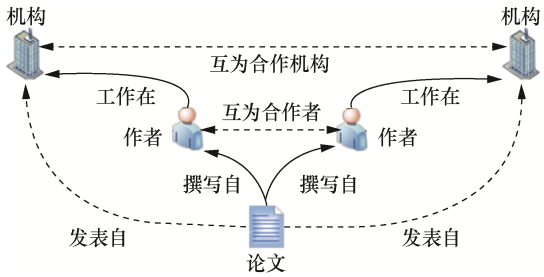



图 2　AceKG 的推理规则[1]

通过上述方式，AceKG 共存储了 6 000 万篇论文、5 000 万名作者、5 万个研究领域、2 万个科研机构、2 万种发表场所以及这些实体间的关系。

该图数据库的优势明显，具体如下。

首先，该图数据库能显著降低查询语句的复杂度，为用户的个性化查询实时生成结果创造了条件。AceKG 以三元组的形式存储知识，当用户进行查询时，AceKG 会对用户查询所用的自然语言进行推理，将其转换为结构化的三元组查询语句，然后将其映射到知识图谱上，从而得到用户查询的结果。如果在传统的关系型数据库中查询，可能需要通过多表联合进行查询，既复杂又浪费时间。

其次，该结构可以构成易于计算机理解的语义网络，在进行推荐任务时，可通过将实体量化后的

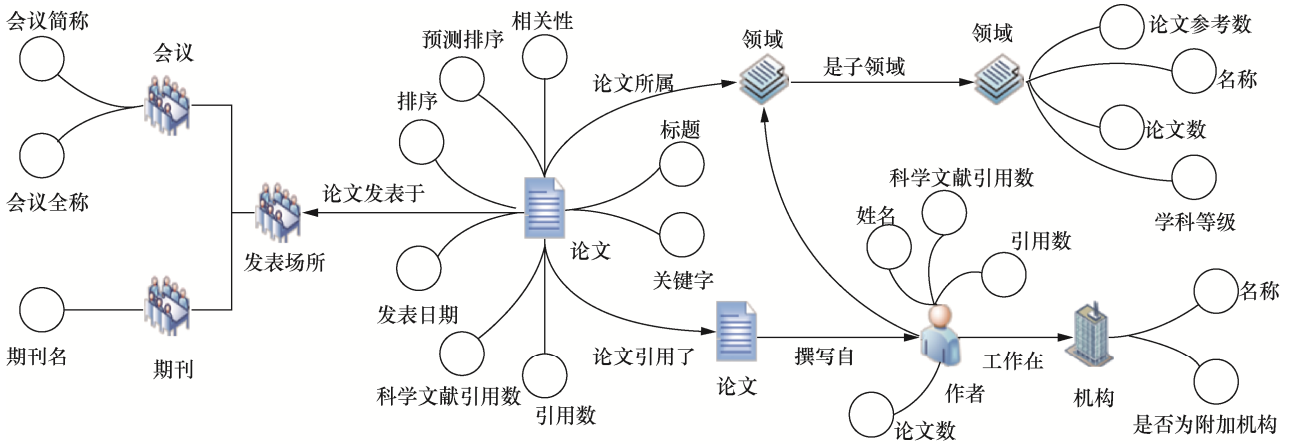



图 1　AceKG 主体结构[1]

结果进行自动推荐，而不需要人为手动设定多种条件，为数据挖掘打下良好基础。首先从已知信息中提取关键词汇，将其与 AceKG 中的实体进行匹配，根据匹配结果从 AceKG 中抽取子图，利用知识图谱特征学习算法学习得到实体向量和关系向量，然后将向量引入推荐模型中进行学习，最终得到推荐结果。因为 AceKG 中包含了大量实体之间的语义关联，可以更深层次地发现用户的研究兴趣，而且 AceKG 中存储的实体间的关系种类丰富，在推荐时可以避免推荐角度过于单一。

## 3 学术地图

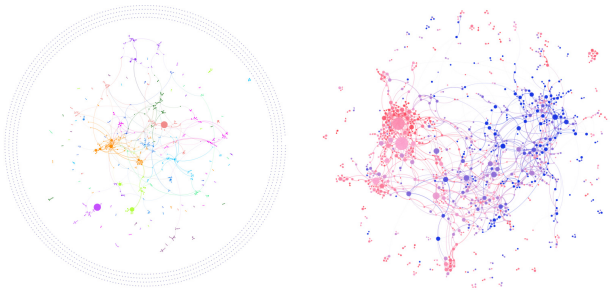
随着物联网概念的延伸，学术数据（如学者、论文、会议以及领域等）逐渐成为其中的节点。近年来，学术发展速度日益加快，物联网中学术数据的规模越来越大。然而，对于用户来讲，不需要了解物联网中全部的学术数据，需要的是某些背景下的数据。因此，为了帮助学者从物联网中获得所需要的数据并挖掘其中隐含的信息，AceMap 通过同构图结构的学术地图对数据进行可视化，帮助用户直观、准确地理解学术领域的发展脉络、学术界的人际关系以及领域之间的异同等。

### 3.1 论文地图

论文是物联网学术数据的重要部分，它可以与物联网中的学者、会议、领域以及其他论文等相连。为了进一步挖掘论文之间的隐含信息，AceMap 将不同学术背景（如会议、领域等）和所有领域下的论文进行了可视化分析，从而帮助用户直观地了解某个背景下论文的影响力大小、不同领域下论文的体量大小以及领域之间的联系。

#### 3.1.1 某一学术背景下的论文地图

当用户需要了解某一会议或领域时，可以直接在 AceMap 上搜索该会议或领域，系统会返回对应的论文地图。图 3(a)为 NIPS 会议上发表的论文地图，图 3(b)展示的是人工智能领域的论文地图。图 3(b)中每个点代表一篇论文，点的大小代表论文影响力的大小，边所连接的两篇论文代表其之间存在引用关系，边的粗细代表引用关系的强度。图 3(a)的外侧有一些孤立的点，说明它们与其他论文不存在引用关系。图 3(b)中点的颜色由红色到蓝色分别代表每篇论文的发表时间由近及远。从图 3(a)、图 3(b)两张地图中可以直观地看出，对其他论文贡献最大的论文、论文之间的相互引用关系和发表时间。对于初次涉猎某个领域的用户来说，可以根据论文地图提供的信息，从最有影响力的论文开始，根据引用关系逐步了解该领域的相关发展。



(a) NIPS 会议　　(b) 人工智能领域

图 3 某一学术背景下的论文地图

#### 3.1.2 全领域论文地图

随着各领域的相互渗透，不同领域的论文在物联网中可以相互联系。图 4 是 AceMap 绘制的全领域论文地图，其中每个点代表一篇论文，不同颜色代表不同领域。从图 4 中可以清楚地看出，每个领域的体量以及不同领域之间的关系，从而对不同领域的发展规模和联系有直观认识。

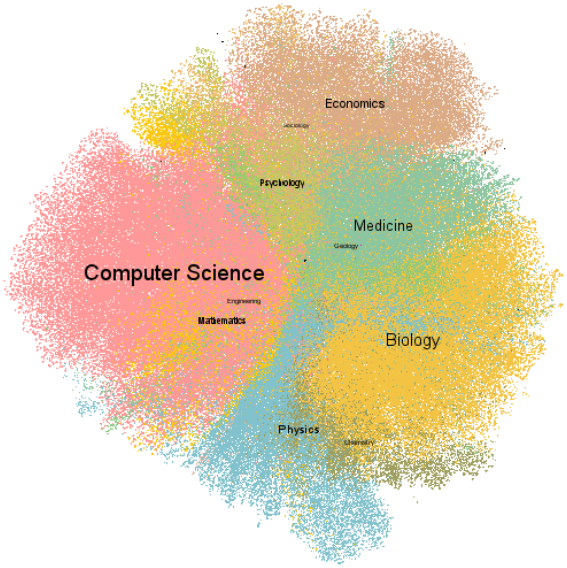



图 4 全领域论文地图

### 3.2 合作者地图

除了论文数据之外，学者数据是学术大数据中非常重要的部分。在物联网中，学者可以与论文、会议、领域以及其他学者相互连接，组成物联网的一部分。为了挖掘学者之间的合作关系和影响力大小，AceMap 对这部分数据中存储的信息进行挖掘，

绘制了合作者地图。

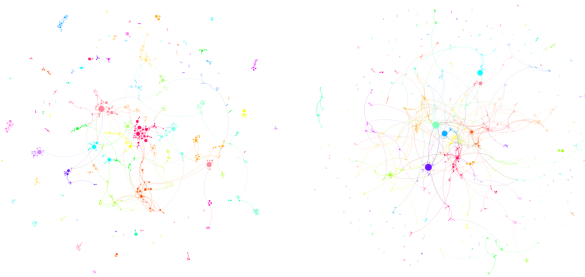
### 3.2.1　以某一学者为中心的合作者地图

在物联网中，以某个学者 *A* 为中心，会有许多其他学者通过合作关系与 *A* 相连，他们之间的合作次数有多有少，合作的影响力也有大有小，AceMap 将这种关系可视化，绘制了如图 5 所示的 *A* 教授的合作者地图。图 5 中每个点代表一个与学者 *A* 合作发表过论文的学者，利用图 5 中每个学者与学者 *A* 合作发表论文的数量以及其他相关数据（如论文的影响力大小），运用 YifanHu[7]算法，可以绘制学者 *A* 的合作者地图。在该地图上，点的大小表示该学者与 *A* 合作发表的论文数目，点越大则说明数目越多；点的颜色表示该作者与学者 *A* 合作论文的影响力大小，点越红则表示该学者与 *A* 发表论文的影响力越大；点与点之间连线的粗细取决于学者间合作关系的紧密程度，如学者 *B*、*C* 之间连线的粗细与 *ABC* 合作论文/min{*AB* 合作论文数，*AC* 合作论文数}成正比。用户从合作者地图中可以清楚地看出，一个学者在学术界的人际关系网络，同时也可以了解哪些学者在论文的编著中起到了重要作用。

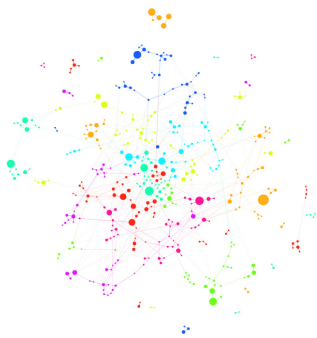

图 5　*A* 教授的合作者地图

### 3.2.2　某一学术背景下的学者合作地图

处于同一个机构、会议或者领域的学者之间合作的可能性更高，在物联网中通过合作关系相互连接的概率更大。为了解某个特定学术背景下学者之间的合作情况，AceMap 绘制了在某一特定学术背景下的合作者地图。在该地图中，展示了用户所搜索学术背景下排名靠前的学者之间的联系。与学者合作者地图不同的是，某一学术背景下的学者合作地图利用 ForceAtlas2[8]算法生成。图 6 分别展示了上海交通大学、NIPS 会议和人工智能领域下的学者合作地图，在图 6 中，每个点代表一位学者，点的大小代表学者发表论文的影响力强弱，点的颜色用以区分不同的学术团体，边代表学者间合作关系的密切程度。以图 6(b)NIPS 会议为例，用户可以直观地看出，该会议中有 4 位学者发表的论文影响力很大，当用户需要了解 NIPS 会议时，就可以先从这 4 位学者的论文入手。除此之外，从点的着色上来看，用户可以清楚地看出学术小团体的数量。当用户想要了解其中某一位学者的学术成果时，除了可以研究该学者的论文，还可以研究图 6(b)中与该学者着色相同的学者论文，因为他们属于同一个学术小团体，意味着其研究方向相近。



(a) 上海交通大学　　(b) NIPS 会议



(c) 人工智能领域

图 6　某一学术背景下的学者合作地图

## 4　结束语

本文介绍了在物联网万物互联的背景下，大量的学术数据成为物联网中的节点。为了能够从物联网中得到更多学术数据的信息，AceMap 设计和完善了学术知识图谱 AceKG，对学术数据进行挖掘，为用户进行个性化查询创造了条件。AceMap 创造了论文地图、合作者地图等不同的地图样式，将所挖掘的信息以同构学术地图的形式进行可视化展示，帮助用户高效获取所需信息。

## 参考文献：


[1] WANG R, YAN Y, WANG J, et al. AceKG: a large-scale knowledge graph for academic data mining[C]//Proceedings of the 27th ACM International Conference on Information and Knowledge Management, October 22-26, 2018, Torino, Italy. New York: ACM, 2018:1487-1490.

[2] SINGHAL A. Introducing the knowledge graph: things, not strings[J]. Official Google Blog, 2012.

[3] HOFFART J, SUCHANEK F M, BERBERICH K, et al. Yago2: a spatially and temporally enhanced knowledge base from Wikipedia[J]. Artificial Intelligence, 2013, 194:28-61.

[4] MITCHELLT M, COHENW W, HRUSCHKAE R, et al. Never-ending learning[C]//Twenty-ninth AAAI Conference on Artificial Intelligence, January 25-30, 2015, Austin, Texas. Palo Alto: AAAI Press, 2015.

[5] JENS L, ROBERT I, MAX J, et al. Dbpedia——a largescale, multilingual knowledge base extracted from Wikipedia[J]. Semantic Web Journal, 2015, 6(2):167-195.

[6] CHRISTOPHER D S, ALEX R, CHRISTOPHER R, et al. Deepdive: declarative knowledge base construction[J]. Sigmod Record, 2016, 45(1):60-67.

[7] HU Y. Efficient, high-quality force-directed graph drawing[J]. Mathematica Journal, 1984, 10(1):37-71.

[8] JACOMY M, VENTURINI T, HEYMANN S, et al. ForceAtlas2, a continuous graph layout algorithm for handy network visualization designed for the gephi software[J]. Plos One, 2014, 9(6).


[作者简介]


**张秋颖**（1995-），女，上海交通大学硕士生，主要研究方向为学术大数据和网络爬虫。

**周乐**（1998-），男，上海交通大学在读，主要研究方向为学术大数据和数据可视化。

**唐静瑶**（1999-），女，上海交通大学在读，主要研究方向为学术大数据和数据可视化。

**傅洛伊**（1987-），女，上海交通大学副教授，主要研究方向为社交网络与大数据。

**王新兵**（1975-），男，上海交通大学特聘教授，主要研究方向为移动互联网、物联网和大数据。