# 《基于用户投诉文本的知识图谱构建研究与实现》文献综述

**1引言**

近年来，大规模知识图谱的研究和应用在学术界和工业界引起了足够的注意力。知识图谱旨在描述现实世界中存在的实体以及实体之间的关系。知识图谱于2012年5月17日由Google正式提出，其初衷是为了提高搜索引擎的能力，改善用户的搜索质量以及搜索体验。在国内，百度和搜狗等搜索公司也相继推出了自己的知识图谱产品。随着人工智能技术的发展和应用，知识图谱逐渐成为关键技术之一，现已被广泛应用于智能搜索、智能问答、个性化推荐、内容分发等领域。

**2知识图谱基本概念**

网络是一个巨大的知识库，其中的知识是并不是孤立存在的，而是相互关联的。换句话说，互联网上的信息是事物而不是字符串[1]从互联网上海量数据中提取相互关联的知识可以构成一个巨大的图。这个图即为知识图谱。知识图谱本质上是一种语义网络。其结点代表实体（entity）或者概念（concept），边代表实体或概念之间的各种语义关系。语义网络[2] 是20世纪50年代末60年代初提出，代表性人物有M. Ross Quillian和Robert F. Simmons。语义网络可以看成是一种用于存储知识的数据结构，即基于图的数据结构，这里的图可以是有向图，也可以是无向图。使用语义网络，可以很方便地将自然语言的句子用图来表达和存储，用于机器翻译[3]、问答系统[4]和自然语言理解[5]。跟早期的语义网络相比，知识图谱具有自己的特点。首先，知识图谱强调的是实体之间的关联，以及实体的属性值，虽然知识图谱中也可以有概念的层次关系，这些关系的数量相比实体之间的关系的数量要少很多，而早期的语义网络主要用于对自然语言的句子做表示；其次,知识图谱的一个重要来源是百科，特别是百科中半结构化的数据抽取得到，这跟早期语义网络主要靠人工构建不一样，通过百科获取高质量知识作为种子知识，然后通过知识挖掘技术可以快速构建大规模、高质量知识图谱；最后，知识图谱的构建强调不同来源知识的融合以及知识的清洗技术，而这些不是早期语义网络关注的重点。

**3知识图谱构建技术研究**

知识图谱的构建可分为三个部分。第一个部分是实体或概念与关系的提取。第二个部分是对提取出来的数据进行融合，除去伪信息和互斥信息，并建立信息之间的关联。第三个部分是以知识图谱为工具进行相关推理。

**3.1实体或概念及关系的提取**

一般而言，实体或概念及关系的提取数据源可分为结构化、半结构化和非结构化三种数据类型。常见的结构化数据有：MusicBrainz，YAGO等。在处理非结构化数据方面，首先要对用户的非结构化数据提取正文。目前的互联网数据存在着大量的广告，正文提取技术希望有效的过滤广告而只保留用户关注的文本内容。当得到正文文本后，需要通过自然语言技术识别文章中的实体，实体识别通常有两种方法，一种是用户本身有一个知识库则可以使用实体链接将文章中可能的候选实体链接到用户的知识库上。另一种是当用户没有知识库则需要使用命名实体识别技术识别文章中的实体。若文章中存在实体的别名或者简称还需要构建实体间的同义词表，这样可以使不同实体具有相同的描述。在识别实体的过程中可能会用到分词[6]、词性标注，以及深度学习模型中需要用到分布式表达如词向量[7]。同时为了得到不同粒度的知识还可能需要提取文中的关键词，获取文章的潜在主题等。当用户获得实体后，则需要关注实体间的关系，我们称为实体关系识别，有些实体关系识别的方法会利用句法结构来帮助确定两个实体间的关系，因此在有些算法中会利用依存分析或者语义解析。如果用户不仅仅想获取实体间的关系，还想获取一个事件的详细内容，那么则需要确定事件的触发词并获取事件相应描述的句子，同时识别事件描述句子中实体对应事件的角色。在处理半结构化数据方面，主要的工作是通过包装器[8]学习半结构化数据的抽取规则。由于半结构化数据具有大量的重复性的结构，因此对数据进行少量的标注，可以让机器学出一定的规则进而在整个站点下使用规则对同类型或者符合某种关系的数据进行抽取。最后当用户的数据存储 在生产系统的数据库中时，需要通过ETL[9]工具对用户生产系统下的数据进行重新组织、清洗、检测最后得到符合用户使用目的数据。

**3.2实体或概念与关系的识别技术**

在知识图谱中，实体与实体或者实体与概念之间存在着关系，关系就是知识图谱上节点之间的边。实体关系识别有流水线方法[10]，还有端到端识别方法[11]。流水线的方法存在的缺点有：1.错误传播，实体识别模块的错误会影响到下面的关系分类性能；2.产生了没必要的冗余信息，由于对识别出来的实体进行两两配对，然后再进行关系分类，那些没有关系的实体对就会带来多余信息，提升错误率。基于统计学的方法将从文本中识别实体间关系的问题转化为分类问题。基于统计学的方法在实体关系识别时需要加入实体关系上下文信息确定实体间的关系，然而基于监督的方法依赖大量的标注数据，因此半监督或者无监督的方法受到了更多关注。无监督学习方面：Bollegala[12]从搜索引擎摘要中获取和聚合抽取模板，将模板聚类后发现由实体对代表的隐含语义关系；Bollegal使用联合聚类(Co-clustering)算法利用关系实例和关系模板的对偶性，提高了关系模板聚类效果，同时使用L1正则化Logistics 回归模型，在关系模板聚类结果中筛选出代表性的抽取模板，使得关系抽取在准确率和召回率上都有所提高。

**3.3实体或概念与关系的融合**

当从上述三种类型数据源中提取实体或概念与关系之后，我们还需要对其进行辨别。因为这些事物可能重复，或者歧义，接下来还要对他们进行融合。最近提出的一种可行的方法使用概率软逻辑（PSL）[13]。它既可以捕获事实之间的概率依赖关系又在实体和其关系之间执行全局约束[14]。在这之后我们会存储这些数据。知识存储和管理的解决方案会根据用户查询场景的不同采用不同的存储架构如NoSQL[15]或者关系数据库。同时大规模的知识 库也符合大数据的特征，因此需要传统的大数据平台如Spark或者Hadoop提供高性能计算能力，支持快速运算。

**3.4知识推理技术**

知识库推理可以粗略地分为基于符号的推理和基于统计的推理。在人工智能的研究中， 基于符号的推理一般是基于经典逻辑（一阶谓词逻辑或者命题逻辑）或者经典逻辑的变异（比如说缺省逻辑）。基于符号的推理可以从一个已有的知识图谱，利用规则，推理出新的实体间关系，还可以对知识图谱进行逻辑的冲突检测。基于统计的方法一般指关系机器学习方法，通过统计规律从知识图谱中学习到新的实体间关系。目前基于统计的方法是主流的选择。其中，基于图特征模型的方法从知识图谱中观察到的三元组的边的特征来预测一条可能的边的存在。典型的方法有基于归纳逻辑程序（ILP）的方法[16]，基于关联规则挖掘（ARM）的方法[17]和路径排序（path ranking）的方法[18]。基于ILP的方法和基于ARM的方法的共同之处在于通过挖掘的方法从知识图谱中抽取一些规则，然后把这些规则应用到知识图谱上，推出新的关系。而路径排序方法则是根据两个实体间连通路径作为特征来判断两个实体是否属于某个关系。

**总结**

在大数据热潮之下，知识图谱也愈发火热了起来。当前知识图谱的研究已经取得了许多成果，但是仍然存在亟待解决的问题。首先是从互联网非结构化的数据中消除噪声问题。其次是在专业领域的知识图谱构建代价较大，因为缺乏相关的文本数据资源。最后就是目前的知识图谱工具实用性较差，推理显得不够智能。

**参考**

[1] Singhal A. Introducing the knowledge graph: things, not strings[J]. Official google blog, 2012.

[2] sowa J F. Principles of semantic networks: Exploration in the representation of Knowledge[J]. Frame Problem in Artificial Intelligence, 1991(2-3):135–157.

[3] simmons r F. technologies for Machine Translation[J]. Future Generation Computer Systems, 1986, 2(2):83-94.

[4] simmons r F. natural language Questionanswering systems: 1969[J]. communications of the ACM, 1970, 13(1):15-30.

[5] Yu Y H, Simmons R F. Truly Parallel Understanding of text[c]// national conference on artificial Intelligence, July 29 - August 3,1990, Boston, Massachusetts, USA.1990:996-1001.

[6]黄昌宁, 赵海. 中文分词十年回顾[J]. 中文信息学报, 2007, 21(3): 8-19.

[7] Mikolov T, Chen K, Corrado G, et al. Efficient estimation of word representations in vector space[J]. arXiv preprint arXiv:1301.3781, 2013.

[8]王敬普, 林亚平, 周顺先, 等. 基于包装器模型的文本信息抽取[D]. , 2006.

[9] 孙安健, 王星, 闫晓瑜. 通用 ETL 工具的研究与实现[J]. 计算机应用与软件, 2012, 29(12): 175-178.

[10] Young D M. Pipelined method and apparatus for processing communication metering data: U.S. Patent 6,377,939[P]. 2002-4-23.

[11]Zhou J, Xu W. End-to-end learning of semantic role labeling using recurrent neural networks[C]//Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). 2015, 1: 1127-1137.

[12] Bollegala D T, Matsuo Y, Ishizuka M. Measuring the similarity between implicit semantic relations from the Web[J]. Www Madrid! track semantic/data Web, 2009:651-660.

[13] Brocheler M, Mihalkova L, Getoor L. Probabilistic similarity logic[J]. arXiv preprint arXiv:1203.3469, 2012.

[14]Pujara J, Miao H, Getoor L, et al. Knowledge graph identification[C]//International Semantic Web Conference. Springer, Berlin, Heidelberg, 2013: 542-557.

[15]Cattell R. Scalable SQL and NoSQL data stores[J]. Acm Sigmod Record, 2011, 39(4): 12-27.

[16] Quinlan J r. learning logical Definitions from Relations[J]. Machine Learning, 1990, 5(3):239-266.

[17] Galárraga L, Teflioudi C, Hose K, et al. Fast Rule Mining in ontological Knowledge bases with aMiE+[J]. The VLDB Journal, 2015, 24(6):707-730.

[18] Lao N, Mitchell T, Cohen W W. Random Walk inference and learning in a large scale Knowledge base[c]// conference on Empirical Methods in natural Language Processing, EMNLP 2011, 27-31 July 2011, John Mcintyre Conference Centre, Edinburgh, Uk, A Meeting of Sigdat, A Special Interest Group of the ACL. DBLP, 2011:529-539.