**基于可度量因果关系网络的常识性因果推理的研究**

摘要

常识性因果推理体现了实现人工智能最本质的挑战，但近年来相关工作进展缓慢。现有知识库中的因果知识都是由人工编撰，数量规模极为有限。人们亟需大规模的常识性因果知识库来帮助实现在常识性因果推理问题上的研究突破。在此背景下，我们提出了一个从互联网海量文本数据中提取并自动化构建因果关系网络的框架，并提出一种新颖的因果度量方法-泛化因果PMI，将因果关系的方向性和强弱程度编码入该网络中。基于此网络，我们设计模型对常识性因果推理问题进行建模。最后，我们将在多种公认、公开的常识性因果推理系统评估数据集上对我们开发的系统进行评测，并实现若干文本应用，以展示我们构建的因果知识库及相关算法的精度及性能。

Abstract

Empowering computers with human commonsense is the long-term goal of AI. Commonsense causal reasoning is the central part of AI challenge, but previous work was not effective due to the lack of large enough commonsense knowledgebase. Commonsense causal knowledge in existing knowledge bases is human-encoded. A well-scaled commonsense causal knowledge base is needed for commonsense causal reasoning task. We propose a framework to automatically harvest a network of causal-effect terms from a web corpus. Based on this network, we propose a novel causal strength metric called generalized causality PMI to encode a causal direction and strength on the network based on lexico-syntactic analysis of sentences. We then modeled commonsense causal reasoning problem by leveraging this network. To demonstrate the usefulness and effectiveness of the proposed causal knowledge and framework, we will evaluate our system on public evaluation dataset and will develop a number of end-to-end applications as well.

二、**报告正文：参照以下提纲撰写，**要求内容翔实、清晰，层次分明，标题突出。

**（一）立项依据与研究内容**（4000-8000字）：

1．**项目的立项依据**

**1.1 研究背景与意义**

常识性因果推理在人们日常行为及决策中扮演着重要角色，它基于对现实生活中各种概念(concept)及其之间的相关关系(relation)的深入理解。同时，它在人工智能领域也是一个十分具有挑战性的问题。

本项目研究的常识性因果推理是主要是对普通知识(常识)所蕴含的因果关系(causality)及因果语义(causal semantics)进行的深入解析。这种解析包括常识性因果关系的检测、识别与提取，并能够在此基础上根据生活中已经发生的事件来预测未来将要发生的事件。因果关系通常由原因(cause)和结果(effect)两个部分组成，在文本数据中通常体现为文本片段(text segment)。常识性的因果检测是识别两个文本片段，如句子或短语之间是否具有因果关系(causality)，例如考虑下面这句话：

“The young campers felt scared **because** their camp counselor told them a ghost story.”

这句话是蕴含了因果关系的，原因和结果都是短语，其中，原因是 ghost story telling, 导致的结果是 feeling scared。

为了进一步阐释这个问题，我们给出了一个因果二选一的问题：

Premise: I knocked on my neighbor’s door. *What happened as an effect?*

Alternative 1: My neighbor invited me in.

Alternative 2: My neighbor left her house.

对于这个问题，显然应该选择第一个选项My neighbor invited me in.在这里，Alternative 1是Premise更可能的结果。我们发现在这个问题中，事件的因果关系位于两个句子，其中隐含原因为knock door，其导致的结果为invite in。当然，考虑knock door这个动作，它还可能导致诸如“refuse to let in”,“not at home”等结果，这就说明同一个因导致的结果是具有多样性的，而invite in则是其中一个概率更高的一个可能。

观察第一个例子，不难发现句子中其实是具有因果提示(causal cue)的，这个提示就是 because， 我们知道如果一个句子中出现了 because 这种词，往往就蕴含有因果关系，这是帮助我们识别因果的重要线索，我们称其为显式因果(explicit causation)。观察第二个例子，我们并没有发现这种因果提示词，但是两个句子所描述的事件之间的确存在因果关系，我们称其为隐式因果(implicit causation)。由此我们不难总结出，在文本数据中因果知识往往以显示(explicit)和隐式(implicit)两种形式蕴含于文本之中[1]。

对于常识性因果推理来说，研究人员亟需构建大规模的常识性因果知识库，但是有如下难点：1) 常识是普通知识，众所周知的知识，人们将其看作是理应知道的一般知识，较少写入文本将其记录下来，而常识性的因果知识则更加稀疏难辨，因而这种知识在通常的文本描述中很难捕捉，这增加了构造常识性因果知识库(commonsense causal knowledgebase)的难度。2) 因果关系是一种较高层次的内在逻辑关系。构建因果知识库涉及对因果语义的深度解析，由于因果知识本身的复杂性及歧义性，因果知识库往往很难保证较高的准确度。

常识性因果推理有着广泛的应用前景，如可应用于问答系统(Question Answering System)，推荐系统，自然语言理解等各领域。例如，常识性推理被用于理解剧本故事[2]。

**1.2 国内外研究现状**

进行常识性因果推理的关键在于提取出常识性因果知识，所以本节我们先讨论目前国内外关于因果关系抽取的相关工作，然后简单介绍关系抽取的一般工作，接着描述研究人员对常识性因果推理这一工作做出的不同尝试，这些工作的基本要素都是单词与单词之间的联系(association)或相似度(similarity)的度量。

**1.2.1 因果关系抽取**

目前，在自然语言处理(natural language processing)领域国内外针对因果关系抽取的工作相对较少 。对于因果知识的识别与提取，传统的研究大致可分为三类，1)基于词项的研究(term-based)，针对词项设计因果特征提取因果知识；2)基于语料的研究(corpus-based)，利用语料设计因果模式或形式化规则提取因果知识；3)基于统计学方法(statistical methods)的研究，利用统计信息度量因果性帮助提取因果知识。

已有方法的一般思路是先进行因果检测，将因果检测问题看做一个二分类问题(causal/non-causal)，设计分类器来进行因果识别，然后再提取出因果对(cause-effect pair)。大多是手工编撰针对特定领域的模式(pattern)来抽取候选因果知识，然后人工设计特征来构建监督学习或非监督学习的分类器，从候选因果信息中识别出真正的因果对。例如，最早进行这项工作的研究人员选择人工设计因果线索和名词特征，构建分类器来抽取名词因果对[1, 3]，在中文方面的相关工作有[4]等。但这些工作并没有充分利用到因果知识出现的上下文信息。另一种思路是认为因果关系是相关性的一种，研究人员设计了用来衡量文本因果相关性的度量来帮助识别文本中蕴含因果关系，如[5]，他们使用因果连接词和相似性分布(similarity distribution)来识别谓词(predicate)之间的因果关系，并取得了0.47的F1-得分。

**1.2.2 一般关系抽取**

除了因果关系的抽取，目前还有很多关于从文本中抽取其他关系的相关工作，如从属(Is-A)关系[6,7]，整体-部分(part-whole)关系[8]，相关(relatedness)关系[9]以及一般(general)关系[10，11,12,13,14,15]等等的抽取。关系抽取涉及到文本中目标词项(target term)和实体(entity)的识别即其关系的标注。

目前的方法主要分为监督学习和半监督学习两种。监督学习方法通常也是把关系抽取看成是一个分类问题，输入是标记了目标词项(target term)或目标实体(target entity)的句子，输出是将其分类到的关系类别，所有关系类别都是事先定义好的。此处用到的实体标记依赖于句法模式或命名实体的识别(Named Entity Recognition)。这些方法要求人工标注的数据作为训练数据，因而不能很容易的扩展到新的关系识别任务中去。并且这些方法需要使用许多重型自然语言处理工具，如词性标注器，依存分析器等等，这些工具本身带来的错误会严重影响在此基础上进行的后续关系抽取工作，导致更低的准确率。半监督学习方法通常设置一个包含若干具有所要抽取关系的实体对作为初始的种子(seed)集合，并采用自迭代的方法通过不断发现特定关系的上下文模式来抽取更多的具有该关系的实体对[6]，另一种选择是事先设置一个固定的集合包含若干特定关系的歧义性较小的上下文模式和逻辑规则，然后用它们对迭代的过程中得到的实体对进行质量评估，以判断其是否具有给定的关系[7]。

**1.2.3 常识性因果推理**

常识性因果推理体现了实现人工智能最本质的挑战。早期对此的一些尝试大多是语言学领域的，例如对因果关系及蕴含关系的形式化理论的研究[16,17]。但由于手工编撰形式化规则的困难与规则适用范围的局限使得这些方法并不是十分有效。近年来，自然语言处理领域的学者已经开发了大量基于知识库的解决领域内各种问题的方法，人们亟需大规模的常识性因果知识库来帮助实现在常识性因果推理问题上的研究突破。ConceptNet[18]是已有的一个具有一定规模的常识性知识库，它自2000年以来招募志愿者为手工其添加常识性知识，里面包含了几种涉及因果性的关系。然而这种手工创建的知识库规模仍受到限制，尤其是其包含的因果知识量十分有限。由于常识性因果知识库的匮乏，多数研究者选择各种无结构化的文本语料作为知识来源，如维基百科文章[19]、古腾堡电子书[20]，微博[21]等等，然后使用语料的统计信息来度量因果性，如点互信息(Pointwise Mutual Information)[22]。我们致力于创建首个大规模自动化构建的常识性因果知识库，将因果知识结构化形成可度量的因果关系网络结构，更加方便使用，并在此基础上提出一套算法以利用该知识库实现常识性因果推理。

目前公认的用来评估常识性因果推理系统质量的数据集也相当有限，被广泛使用的主要有RTE (Recognizing Textual Entailment)数据集[23]，及在它的基础上构造的COPA(Choice Of Plausible Alternatives)评估数据集[20]。RTE数据集包含了1600个判断问题，每个问题给定一段文本T(text)和另一段文本H(hypothesis)，来判断两段文本间是否具有蕴含关系(entailment),例如如下这道判断题要求系统判别T与H之间是否存在蕴含关系：

T: Cavern Club sessions paid the Beatles £15 evenings and £5 lunchtime.

H: The Beatles performed at Cavern Club at lunchtime.

COPA数据集是一个包含了1000个常识性因果推理选择题的数据集，包括已知原因推结果和已知结果推原因两种类型的问题，例如：

Premise: I knocked on my neighbor’s door. What happened as an effect?

Alternative 1: My neighbor invited me in.

Alternative 2: My neighbor left her house.

这是一道已知原因问结果的问题，再如：

Premise: The man fell unconscious. What was the cause?

Alternative 1: The assailant struck the man in the head.

Alternative 2: The assailant took the man’s wallet.

这是一道已知结果问原因的问题。我们将使用这些公认的数据集来评估我们开发的常识性因果推断系统。同时我们也可以通过人工标注形成一些有影响力的自主数据集。

**1.3参考文献**

[1] R. Girju. Automatic detection of causal relations for question answering. In Proceedings of the ACL 2003 workshop on Multilingual summarization and question answering, pages 76–83, 2003.

[2] MUELLER, E. T. Understanding script-based stories using common- sense reasoning. Cognitive Systems Research 5, 4 (2004), 307–340.

[3] D. Chang and K. Choi. Causal relation extraction using cue phrase and lexical pair probabilities. In IJCNLP, pages 61–70, 2004.

[4] 丁小杉. 基于句子的因果关系识别. 上海交通大学, 2010.

[5] Q. X. Do, Y. S. Chan, and D. Roth. Minimally supervised event causality identification. In EMNLP, pages 294–303, 2011.

[6] O. Etzioni, M. Cafarella, D. Downey, S. Kok, A. M. Popescu, T. Shaked, S. Soderland, D. S. Weld, and A. Yates. Web-scale information extraction in knowitall. In WWW’04, 2004.

[7] W. Wu, H. Li, H. Wang, and K. Q. Zhu. Probase: a probabilistic taxonomy for text understanding. In SIGMOD’12, 2012.

[8] R. Girju, A. Badulescu, and D. I. Moldovan. Automatic discovery of part-whole relations. Computational Linguistics, 32(1):83–135, 2006.

[9] K. Q. Z. Keyang Zhang and S.-W. Hwang. An association network for computing semantic relatedness. In AAAI, 2015.

[10] A. Yates, M. Cafarella, M. Banko, O. Etzioni, M. Broadhead, and S. Soderland. Textrunner: Open information extraction on the web. In NAACL HLT Demonstration Program, 2007.

[11] 车万翔, 刘挺, 李生. 实体关系自动抽取. 中文信息学报, 2005, 19(2): 1-6.

[12] F. M. Suchanek, G. Kasneci, and G. Weikum. YAGO: a core of the semantic knowledge. In WWW’07, 2007.

[13] J. Hoffart, F. M. Suchanek, K. Berberich, E. L. Kelham, G. de Melo, and G. Weikum. YAGO2: Exploring and querying world knowledge in time, space, context, and many languages. In WWW’11, 2011.

[14] A. Fader, S. Soderland, and O. Etzioni. Identifying relations for open information extraction. In EMNLP, pages 1535–1545, 2011.

[15] N. Nakashole, G. Weikum, and F. M. Suchanek. PATTY: A taxonomy of relational patterns with semantic types. In EMNLP-CoNLL, pages 1135–1145, 2012.

[16] A. Lascarides, N. Asher, and J. Oberlander. Interferring discourse relations in context. In ACL, pages 1–8, 1992.

[17] A. Lascarides and N. Asher. Temporal interpretation, discourse relations and commonsense entailment. Linguistics and Philosophy, 16(5):437–493, 1993.

[18] H. Liu and P. Singh. Commonsense reasoning in and over natural language. In Knowledge-based intelligent information and engineering systems, pages 293–306. Springer, 2004.

[19] S. Jabeen. Exploiting wikipedia semantics for computing word associations. 2014.

[20] M. Roemmele, C. A. Bejan, and A. S. Gordon. Choice of plausible alternatives: An evaluation of commonsense causal reasoning. In AAAI Spring Symposium: Logical Formalizations of Commonsense Reasoning, 2011.

[21] A. S. Gordon, C. A. Bejan, and K. Sagae. Commonsense causal reasoning using millions of personal stories. In AAAI, 2011.

[22] A. S. Gordon, Z. Kozareva, and M. Roemmele. Semeval-2012 task 7: Choice of plausible alternatives: An evaluation of commonsense causal reasoning. In the 6th International Workshop on Semantic Evaluation (SemEval 2012), 2012.

[23] RTE : <http://pascallin.ecs.soton.ac.uk/Challenges/RTE/>

2．**项目的研究内容、研究目标, 以及拟解决的关键科学问题**（此部分为重点阐述内容）**；**

**2.1 研究目标**

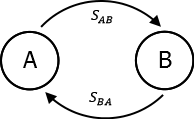
人们常常提到“因果关系”这一词，但都没有给予它一个明确的定义，我们首先定义因果在本项目研究中的概念：如果一件事(事件)发生的原因是另一件事(事件)，我们就说这两者之间存在依赖或因果关系。如图1，A为因，B为果，A到B组成了一对因果事件。



**图1 因果关系对示意图**

对于因果关系，我们需要设计模型来解决如下问题：

我们如何描述事件之间的因果关系，是使用概率，可能性，置信度还是定义一种因果关系强度。这种因果关系强度应该能够表示事件之间相互因果关系的强弱，如图2所示：



**图2 两个事件之间的相互因果关系**

其中，。

我们将对大量因果事件构建一个具有这样因果关系强度的大规模因果网络，作为常识性因果关系知识库，我们称其为CausalNet。然后设计算法利用CausalNet，来解决常识性因果推理问题。

常识性因果推理系统的目的是推断出日常语言中会出现的命题(proposition)之间是否具有因果关系。我们考虑COPA数据集中出现的两种常见类型的因果推理问题，1) 给定原因推出结果，2)给定结果推出原因，其中原因(cause)和结果(effect)均为文本的描述。

本项目目旨在通过对常识性因果推理模型的研究，设计并实现一套完整的算法与原型系统。包括从海量互联网网页文本数据抓取一般性因果关系知识，利用其统计信息自动构建首个基于单词和事件的大规模可度量因果关系网络(CausalNet)，并基于该因果关系网络对常识性因果推理问题进行建模，设计并实现常识性因果推理算法，有效提高推理系统的准确度。CausalNet可以克服目前由手工编撰常识性知识网络产生的知识稀疏，耗费大量人力财力，更新效率低，速度慢等局限性，从而以更低成本，更高的准确度更好地解决常识性因果推理问题。

**2.2 研究内容**

本项目的主要研究内容是在名词短语知识库和语义角色标注的工作基础上，

1）分析不同文本语料对常识性因果知识的贡献并充分利用这些语料；

2）设计与实现从海量文本语料中抽取一般性因果知识的算法，抓取出大量存在于单词(word)之间及事件(event)之间的因果关系信息；

3）设计和实现对抽取数据清洗和去噪；

4）建立因果关系模型，构建出可度量化的因果关系网络(CausalNet)；

5）设计与实现常识性因果推理算法；

6）将我们的因果推理系统应用到如下工作中，以评估系统的有效性， a) COPA数据集，b) 文本隐含关系识别c)问答系统。

下面我们对上述几个方面的研究做简单阐述。

**2.2.1 分析语料**

我们首先分析现有语料、知识库对常识性因果推理工作的贡献及意义，并充分利用这些语料识库来协助构建我们的常识性因果推理系统。最简单的知识库就是词典(lexicon)，如目前应用最为广泛的WordNet词典。WordNet中的基本单元是同义词集(synset)。WordNet可以告诉我们每个单词对应到不同同义词集下的解释(glossary)，还能帮我们找到与给定同义词集具有各种相关关系的其他同义词集，包括同义词、近义词、反义词等，其中对于因果推理工作最重要的关联关系是因果关系(CAUSE)，例如“ferment CAUSE sour”，“strain CAUSE tense up”。除此之外，蕴含关系(ENTAILMENT)也相当重要。我们可以把蕴含关系简单的理解成一件事是另一件事的前提条件，例如“succeed ENTAILMENT attempt”，“wake up ENTAILMENT sleep”等等。这里sleep是wake up的前提条件。另外一种颇具规模的常识性知识库是麻省理工大学发起创建的ConceptNet，ConceptNet具有几种与因果关系相关的关系如“Causes”，“CausesDesire”和“HasPrerequisite”。例如“rain ***-Causes->*** make grand wet”这条因果知识可以由人工录入的“The effect of rain is making the ground wet.”这条常识得到。ConceptNet也是人工创建的，至目前为止已耗时15年，但规模仍然有限，尤其是其因果关系知识含量较小。

此外，由于因果关系和蕴含关系是较高层次的语义关系，具有相当的复杂性和歧义性，例如我们不能简单的说抽烟(smoking)与死亡(death)之间是否具有因果关系，在不同的情况下抽烟对于发生死亡这一事件的影响程度是不同的，但如果利用人们的常识来判断这两者之间应该是具有一定的因果关系的。由于上述因素因而编撰这些知识十分困难，上述这些语料中也只含有极少量的因果知识，并且这些因果知识并没有一个量化的标准来衡量原因与结果之间的因果强弱，这远远不能满足让机器进行常识性因果推理的需要。但我们可以合理利用这些信息可以帮助我们识别和提取因果关系知识。

**2.2.2 候选因果知识抽取算法**

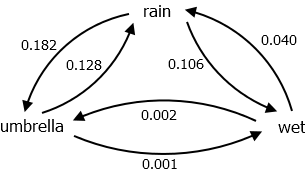
我们利用互联网的海量文本数据作为基础语料库，利用因果线索模式从中提取显式因果知识作为候选因果知识。显式因果知识具有准确率高的优点，但是相应的数量较少，但是由于我们的基础语料巨大(约为10TB量级)，得到的候选知识数量仍然相当可观。因为处理对象是巨大文本语料，我们选择轻量级处理，简单的根据因果模式事先定义好的原因和结果角色直接进行抽取，并没有使用词性标注器，依存分析器等重量级的自然语言处理工具。

**2.2.3 候选数据去噪**

在从海量数据中提取因果词项/事件对时，为了加速处理过程、减少时间和空间的损耗，我们使用较为简单的词项/事件提取的方法。然而，这种方法也给我们的数据带来了一定的噪声。首先，不论是句间还是句内的因果关系，词的词性以及在句中的成分，都对因果关系有不同的贡献。例如，名词和动词比之副词更容易成为因或果；动词和宾语比之状语更容易成为因或果。而在提取过程中则忽略了这种句内结构。另外，有一些高频常用词，本身对因果关系并不作出任何贡献，例如be动词，限定词等，我们把它们统称为停用词（stop words），也会对我们作出判断造成干扰。由此可见，我们需要设计一些方法对原始数据进行适当的清洗和去噪，过滤掉无用词，削弱效用低的词的作用，强化有效词的意义，以提高后续计算的准确度。

**2.2.4 建立可度量化的因果关系网络CausalNet**

在对抽取的数据进行清洗去噪之后，我们得到了大量且较为准确的因果词对或因果事件对，实现了从非结构化到半结构化因果知识的成功转化。接下来，要设计可度量因果性的方法，将半结构化的因果知识转化为结构化的因果知识，即创建可度量化的因果关系网络CausalNet。为说明这一想法，我们给出CausalNet网络的一个具有三个节点的子图，这里我们将事件简化成词项：



**图3 CausalNet子图示意图**

由图3可以看出单词rain导致单词wet的因果强度理应强于单词wet导致单词rain的因果强度，因为下雨导致潮湿是显而易见的，然而导致潮湿的原因却有很多，并不一定是下雨。当然，词项并不能囊括所有的事件，事件通常不是一个词而是由一个文本片段的形式出现，我们同样需要设计模型来描述事件。我们将在3.2.3节讨论因果性度量的方法设计。

**2.2.5常识性因果推理算法**

有了可度量的常识性因果关系网络作为我们的因果知识库，我们就可以利用这个知识库设计因果推理算法来解决常识性因果推理问题。我们设计了几种不同模型来使用CausalNet。思路是将原因中的词项/事件看做是触发器(trigger)，将结果中得词项/事件看作是响应(response)，触发器与响应之间可以被建模成一对多和多对多等关系。我们将在3.2.4节中详细描述常识性因果算法的设计。

**2.2.6 文本处理示范应用**

本项目将实现下列几个文本处理中常见的重要应用：

1) 对于给定原因/结果,对其相应的若干结果/原因进行合理性排序，可利用COPA数据集进行评估。比如在Alternative 1和Alternative 2中选择一个更合理的答案：

* Premise: The man lost his balance on the ladder. What happened as a result?
* Alternative 1: He fell off the ladder.
* Alternative 2: He climbed up the ladder.

2) 判断文本之间存在的蕴含关系。比如判断如下两句话之间是否存在蕴含关系：

* T: Cavern Club sessions paid the Beatles £15 evenings and £5 lunchtime.
* H: The Beatles performed at Cavern Club at lunchtime.

3）问答(QA, Question Answering System)系统，对于给定的剧本故事，能够帮助问答系统理解文本并回答更深层次的问题，例如给出与恐怖主义相关的故事文本之后，可以利用从中解析出的因果推理模板回答以下原文中并没有显式提到的问题：

* Q: Who threatened the villagers?
* A: The unidentified persons.
* Q: Who grabbed the villagers?
* A: The unidentified persons.
* Q: What did the unidentified persons set fire to?
* A: The grocery store.

**2.3 拟解决的关键科学问题**

要达到上述的研究目标，本项目必须解决以下几个关键科学问题。

1. 找到最合适提取常识性因果关系的语料。由于因果知识的稀疏性，我们必须选用大规模基础预料抽取因果知识，如海量互联网文本数据，目前我们具有的文本快照数据量约为10TB。另一种可能是从因果关系比较密集的文本，如叙述性强的故事中提取。此外，我们还充分利用已有的语料库对因果知识进行的已有解析处理工作，如WordNet，ConceptNet等。
2. 构造大规模常识性因果关系知识库。从海量文本数据中抽取大量候选因果知识，并利用机器学习算法设计分类器去除因果关系图中的噪声。
3. 构建因果关系图结构并利用抽取数据的统计信息使因果关系网络可度量化。
4. 设计算法模型来计算文本片段之间的因果关系强度。

3．**拟采取的研究方案及可行性分析**（包括研究方法、技术路线、实验手段、关键技术等说明）；

**3.1研究方法**

本项目的研究主要采用文献系统调研、算法设计与实现、实验验证评估然后继续对算法及系统实现进行微调这一渐进式研究方法（见图4）。项目的推进大致按照图中箭头所指的方向，由于整个项目的周期比较长，在研究过程中可能会出现新的相关工作和研究成果，但上述过程并非完全按照顺序进行，而是彼此穿插，在循环迭代中前进，直到达到所有评测指标为止。



**图4 本项目研究方案**

下面结合图4简单介绍一下本项目研究方案中的三个重要环节。

1. 在调研环节，我们首先详细阅读包括知识库构建、文本数据理解、自然语言解析方法、关系抽取在内的相关文献资料。然后对本项目所需的工具、系统（如Stanford的自然语言处理工具集，WordNet、ConceptNet等常识性知识库），进行评测，优化配置资源。同时我们将从微软Bing检索的万维网网页快照中抽取所有包含特定因果模式的词对/事件对。这里我们要利用机器学习方法做必要的预处理，筛除噪声，得到精度较高的半结构化的常识性因果知识。
2. 在算法设计实现环节，我们会为每个模块设计一种主要方法和至少一种备用方法（可以是主要方法的变种）。如果主要方法经过初步评估无法达到预期效果，我们会实现备用方法，对两者进行比较。

3）在评估环节，我们会对系统和算法做两种评估。首先是评测算法对已有语料中存在的因果知识的推理辨别能力。我们将从ConceptNet中提取人工标注出来的正确和错误因果关系知识作为标准数据集，用其对我们的系统进行检验。其次是利用已有的公认的常识性因果推理评估数据集，如COPA等等，来评估我们的系统在常识性因果推理问题上的表现。这里大部分的应用都存在可以比较的现有系统或算法，我们将采用现成系统或实现一些对比算法然后在公开或公平的数据集上做对比测试。

**3.2 技术路线**

下面我们对算法中几个主要模块以及几个关键应用的技术路线做简单设想。

**3.2.1 因果关系对的抽取**

文本中蕴含的因果关系可以分为显式（explicit）因果关系和隐式（implicit）因果关系两种。为便于解释说明，考虑下面几个句子：

(1) The *storm* **caused** a tremendous amount of *damage* on the landing beaches.

CAUSE PATTERN EFFECT

(2) The team prepared GIS precipitation and contour maps of the area identifying the *flooding* and landslides **caused by** the *rainfall*.

EFFECT PATTERN CAUSE

(3) The *discomfort* from the *injury* was now precluding him from his

EFFECT CAUSE

occupation which involved prolonged procedures in the standing position.

文本中的显示因果关系通过句中或句间出现的因果模式（causal patterns）来表示，如(1)和(2)所示。同时，文本中也蕴含了大量隐式的因果关系，如(3)所示。可以看出，由于隐式因果关系出现在文本中并没有固定的模式，识别它们要困难的多，往往会引入大量的噪音（noise）。因此，我们期望从显示因果关系中提取因果知识，以保证因果知识的准确性（即质量）。同时，我们还需要权衡得到因果知识的覆盖率（即数量）。这种权衡主要体现在因果模式的设计上。我们初步设计了53种因果模式,包括句内因果模式（intra-sentence causal pattern）和句间因果模式（inter-sentence causal pattern），如表1所示。并且，我们使用了大量的文本语料（10T左右）以便能够覆盖到足够多的因果信息。值得注意的是，因果模式可以更加准确的确定文本中因果关系‘因’(cause)、‘果’（effect）角色（role）。在从文本中提取因果关系时，表1中‘A’部分的单词标记为‘因’，‘B’部分的单词标记‘果’。

**表1： 因果线索模式**

|  |  |  |  |
| --- | --- | --- | --- |
| intra-sentence | | inter-sentence | |
| A lead to B | A leads to B | If A, then B | Inasmuch as A, B |
| A leading to B | A give rise to B | B because A | B because of A |
| A given rise to B | A giving rise to B | A, thus B | A, therefore B |
| A inducing B | A induces B | If A, B | B, inasmuch as A |
| A cause B | A causing B | B due to A | A and consequently B |
| A caused B | B caused by A | B owing to A | B as a result of A |
| A brought on B | A bringing on B | A and hence B | Owing to A, B |
| B result from A | B resulting from A | A, hence B | A, consequently B |
| A led to B | B resulted from A | Due to A, B | B in consequence of A |
| A gave rise to B | A bring on B | B, because A | As a consequence of A, B |
| A induce B | DET effect of A BE B | Because A, B | B as a consequence of A |
| A induced B | B results from A | B, A as a consequence | |
| A causes B | the reason(s) for/of B BE A | In consequence of A, B | |
| A brings on B | A BE DET reason(s) of/for B | A, for this reason alone , B | |

其中，DET表示限定词，例如：a、an、the、one. BE表示be动词，例如is、are、was、were。

我们从这些含有因果模式的文本中提取因果知识，即因果关系词对（causal pairs），我们采用的方法如下：

检索句子中包含有关键词的句子，将A中的单词作为‘因’（记为）与B中的单词作为‘果’（记为）两两相连成因果对，记其出现频率为1，同时记录下包含的相应关键词。将所有的频率相加，我们得到不同的词对对于不同的关键词的频率分布以及它的总频率。例如，从句(1)中，可以提取出的因果关系对有：，，，和。在此过程中，我们去除了停用词（stop word），并且只保留在WordNet中作为名词、动词、形容词和副词出现过的单词。

**3.2.2 数据清洗与去噪**

由于句子结构和停用词等的存在，原始数据中存在一定的噪声。我们将从以下几个角度进行数据清洗和降噪：

1. 构造停用词词典。对于任一因果词（事件）对，如果包含了停用词，则忽略词（事件）对，不纳入计数。拟采用的停用词词典部分表2所示：

**表2：停用词表**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| a | afterwards | alone | among | any | anywhere | av |
| about | again | along | amongst | anybody | apart | be |
| above | against | already | am | anyhow | are | became |
| according | albeit | also | an | anyone | around | because |
| across | all | although | and | anything | as | become |
| after | almost | always | another | anyway | at | becomes |

2）对每一个单词进行正则匹配，过滤非法字符。

3）统计词（事件）对的频率，考查其分布，选择合适的阈值，滤去低频不常用词。

4）从提取的词（事件）对中随机挑选出一定数量的样本，人工标注其正负因果，尝试采用训练出不用的分类器：神经网络、支持向量机SVM、深度学习等方法，对原始数据分类，保留具有因果关系的词（事件）对。

5）根据人工标注的结果，评估不同pattern对于其因果关系的影响，调整分数。

**3.2.3 因果关系度量化模型**

目前有工作使用基于词汇共现次数（lexical co-occurrence）的点互信息(PMI, Pointwise Mutual Information)作为因果关系强度度量。在典型的叙事文本语料（如个人故事，personal story）中，事件通常是按照时间顺序发生及叙述的。因而，文本中先出现的事件更可能是原因，而后出现的事件更可能是它的结果。这使得使用词汇共现次数来计算PMI更加合理。由于不需要因果模式，提取的是文本中的隐式因果关系，覆盖率也得到保障。传统的点互信息定义如式1，此处u和v为单词：

当我们使用PMI来作为度量为因果对排序时可以省略掉其对数形式，得到如下形式，如式2：

还有其他的一些备选度量方法，如Dice系数，Jaccard指数等等。

但是，这种方法要结合特定的叙事文本进行使用，否则会引入大量噪音。而这种叙事性文本不易得到，需要耗费大量的人力进行收集。我们提出了通过因果模式从大量web文本语料中提取因果信息的方法。由于可以利用更易获取的web语料，提取的因果信息规模得到了保证。此外，我们设计的因果模式可以更准确的标识出因果关系中的原因和结果。这使得我们提取的因果关系对更加准确。

如上,通过因果模式引入了因果角色信息之后，我们可以对现有因果度量计算方法进行改良。本质上，是用显式的因果共现（causality co-occurrence）频率取代隐式的词汇共现（lexical co-occurrence）频率。我们的观点是，因果关系主要表现为充分性因果（sufficiency causality）和必要性因果（necessity causality）的结合。考虑因果关系对，充分性因果定义为单词作为‘因’（即）一定会导致‘果’（即）；而必要性因果定义为作为‘果’发生，那么一定是作为它的‘因’发生了。当然，在实际当中一个因果关系对通常都不是纯粹的充分性因果或必要性因果，而是二者的结合。例如，句(2)中因果对就更倾向于必要性因果，因为如果‘因’rainfall没有发生，则‘果’flooding通常不会发生。同样的，则是更倾向于充分性因果的关系对。我们对这两种因果关系做适当的描述及建模，并由此提出了种新的度量方法来计算因果强度。直观上来说，必要性因果越强则越大；而充分性因果越强则越大。此外，我们还对高频词做了惩罚。得到了式3所示的充分性因果强度和必要性因果强度的计算公式：

其中,是指数惩罚常数。

M是所有因果关系的总数，N是文本语料的大小。W是语料中所有单词的集合。通过设计上述因果关系度量方法，我们可以度量化得到的因果关系知识，形成带权因果关系网络CausalNet。

**3.2.4 常识性因果推理算法设计**

这部分是我们利用CausalNet进行常识性因果推断任务的关键。通过因果PMI的设计，我们很容易得到词项/事件之间的因果性强弱分数，我们提出几种利用词项/事件之间的因果性计算任意两个文本片段A(原因)，B(结果)之间因果性的算法。同时，我们还可以利用词项在WordNet中所属同义词集数目的多少等信息，来削减歧义词的过高因果性分数。

1. **全匹配(all pairs)算法**

将词项及事件作为文本片段基本单元，将A中所有单元与B中所有单元两两匹配利用CausalNet得到其因果性，并求和作为A、B之间的因果性强弱的分数，如式5：

其中,, 表示在WordNet中u所在的同义词集数。

1. **最大化(max)算法**

将A中的每个基本单元与B中所有单元比配留下因果分数最大的那个匹配对，将所有留下的匹配对的因果分数相加，作为A、B之间的因果性强弱的分数，如式6：

1. **阈值(threshold)法**

设计因果分数阈值，留下大于阈值的匹配对，将其分数求和作为A、B之间的因果性强弱的分数，如式7：

记为

令

其中t为阈值。

**3.3 实验验证**

我们将严格遵守在3.1节中评估环节的步骤，利用已有的公开、公认的常识性因果推理评估数据集来评估我们的算法，并采用现成系统或实现一些对比算法进行对比测试。

**3.4 项目可行性分析**

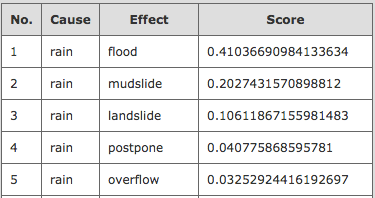
本项目的可行性由以下几个方面来保障：

首先，我们使用海量互联网文本数据作为基础语料，使得自动化构建大规模因果知识库称为可能。

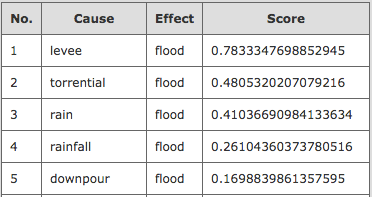
其次，常识性因果推理相关的公开数据集的提出，使得相关工作的评估更加容易，将会推动因果推理相关研究领域的发展。

第三，我们在前期研究中已经初步实现了词项对之间因果性度量的一些工作，在这里可以提供一些演示示例：

示例1，rain 作为原因，排名在前五的结果及其因果性分数如下图所示：



示例2，flood作为结果，排名在前五的原因及其因果性分数如下图所示：



第四，本项目的依托单位教育部-微软智能计算重点实验室配备多台高性能计算机以及计算集群，保证项目组的强大的计算能力和充足的储存空间。

最后，项目申请人和微软亚洲研究院的紧密合作关系保证了我们可以真正从大数据上构建该知识库并获得真实的实验数据（如Bing检索的网页和储存的搜索日志等），同时我们在大数据上的实验验证也可以在微软的超大规模云计算平台Cosmos上实现。申请人和普林斯顿大学、新加坡国立大学、香港科技大学、AT&T 研究所、Google、 Yahoo、IBM研究院等著名大学和研究机构也保持这广泛合作和联系。这保证本项目的国际视野，使成果更容易被国际同行接受。

4．**本项目的特色与创新之处**

首先，本项目使用文本大数据解决人工智能领域的重大挑战。本项目所研究自动构建的常识性因果知识库CausalNet，是常识性因果推理领域内的一项创举。目前还没有一个专门为因果关系自动构建的大规模因果知识库。现有知识库（WordNet、ConceptNet等）中的因果知识都是由人工编撰，数量极为有限。无论是在规模、成本还是专业性和可扩展性上我们的CausalNet都胜过已有知识库。

其次，我们将图算法和自迭代算法相结合，来去除候选因果关系数据中的噪声，在保证规模的前提下同时提高了因果知识的质量。本项目建立的因果关系网络的精确度和覆盖面都是前所未有的。

第三，本项目的另一贡献是定义了度量因果性强弱的方法，泛化因果PMI，并设计因果关系推理算法，从而实现可以为任意两个文本片段计算因果性分数。

第四，本项目基于CausalNet设计常识性因果推理系统是领域内工作的一种全新的尝试，之前的方法大多基于非结构化的文本语料，并未得到结构化的因果知识库，我们研究开发的是更复杂更细化的因果知识。这一工作对于文本理解、文本检索和问答系统等一系列应用有直接驱动作用。

综上所述，本项目借助自动构建的常识性因果知识库CausalNet设计一系列算法实现了常识性因果推理系统。把文本因果推理和因果语义分析研究带到了一个新的高度。此项目的成功实施预示着因果关系等高级语义关系正在文本理解及推理领域扮演着的越来越重要的角色。常识性因果推理不仅帮助我们揭开文本中暗藏的深层语义信息，也向自然语言处理和高性能计算提出了新的挑战。

5．**年度研究计划及预期研究结果**（包括拟组织的重要学术交流活动、国际合作与交流计划等）。

**研究计划**

2015.1 – 2015.6 文献调研，采集实验数据。

2015.7 – 2015.12系统和资源调研和测试。实验数据预处理。

2016.1 – 2016.6 实现因果知识的抽取模块和数据去噪模块及相关算法。

2016.7 – 2016.12 实现因果知识库的构建和因果性度量方法，并实现常识性因果推理相关算法。

2017.1 – 2017.6 大数据上评估算法和系统，对比其他算法，微调系统。

2017.7 – 2017.12 设计实现3-5个重要应用并与现有算法比较。

2018.1 – 2018.6 实现网上Demo，申请专利，推广项目成果。

2018.7 – 2018.12撰写论文与结题报告。

另外我们还会和微软亚洲研究院的数据库组做至少半年一次的定期的互访，并举办学术交流会。本项目的研究生将被派驻研究院，完成一些大规模或真实数据上的实验验证。

**预期研究成果**

1. 建立一个大规模的、全方位的带权常识性因果知识库
2. 建立一套完整的常识性因果推理算法及框架
3. 实现至少三个重要文本处理应用，并展示其相对传统自然语言算法在性能及精度上的显著提升
4. 在国际顶尖会议和期刊上发表高水平学术论文6－10篇
5. 申请至少三项国际或国内技术发明专利

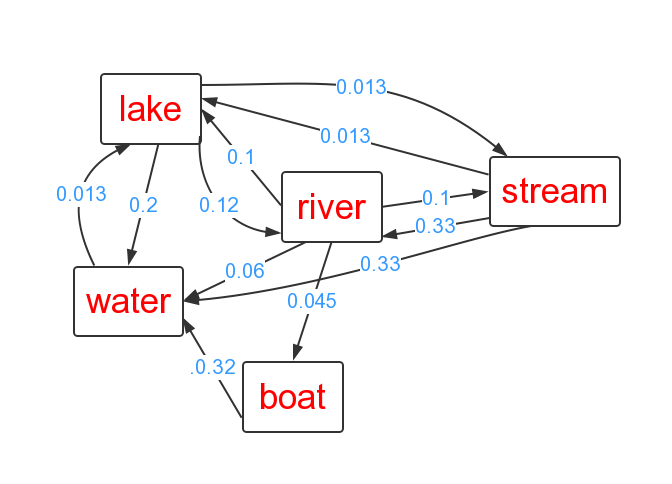
**（二）研究基础与工作条件**

**1．工作基础**（与本项目相关的研究工作积累和已取得的研究工作成绩）；

本项目申请人多年来从事文本数据分析处理、信息抽取与知识发现方面的研究工作，取得一定成果。下面从四个方面概括与本项目相关的工作积累和成绩。

* 1. **联想网络与文本相关性**

受到自然科学基金中韩合作交流基金的资助，我们在2014年与韩国POSTECH的Seung-won Hwang教授合作，从维基百科的文本以及链接结构中提取出一个大型的概念联想网络（如下图），这个网络可以模拟人一种称为自由联想的人脑认知活动，比如人们一般会从“河流”联想到“湖泊”、“轮船”、“水”等等。



该网络包含数万个常用英文词汇的节点，以及它们之间的单向关联度。我们在这个网络上设计了词语之间以及短句之间相关度的计算方法，并在包括WS-353、Li-30等业界公认的标准测试集上进行评测，对于词语间相似度，我们达到0.813准确度，在短句相关度上达到0.94的准确度。这两个结果均优于目前最先进的相关度算法。相关工作以长文发表在AAAI 2015上，并申请了国家发明专利一项。本项目中研究的因果关系是词语间联想关系的一个特例，因此我们对联想关系的抽取和词语相关度计算的研究可以对本项目产生指导作用。

* 1. **动作概念化**

我们在万维网网页数据以及Google Syntactic N-gram中提取了近一万个英文动词以及它们的主语和宾语。然后我们利用Probase或WordNet中的isA关系，将同一动词下的主语和宾语抽象成若干概念，例如对于动词Play,我们可以将其抽象成Player-PLAY-Sports, Musician-PLAY-Musical\_instrument等*k*个“动作概念”。事实上这种为动词的参数(argument)概念化的方法可以被视为一种粒度可调的（调节k）Word Embedding的方式。这样，大量的动作概念三元组可以组合成一个动作概念图（如下图），作为我们理解动词在语句中语义的一种有效手段。这项工作获得2013年度Google Faculty Research Award，也是当年中国大陆唯一的获奖者。在本项目中，抽取语句中的事件(events)需要有效地识别出动词和它的参数，而参数判别正是动作概念化的一个具体应用。



* 1. **大规模从属关系(isA)概念知识库**

2010年－2013年，申请人与微软亚洲研究院数据库组合作研究，利用云计算和智能数据挖掘技术从万维网数据（16.8亿网页）上自动构建了迄今为止最大规模（包含超过280万个概念）的基于概率模型的从属关系(isA)概念知识库Probase（发表在SIGMOD 2012，Google Scholar统计被引用超过140次)。Probase是完全自动从海量的万维网数据中抽取构建出的分类知识库(taxonomy)。它将各种概念和实体之间的从属关系在一个巨大的图结构中表达出来。图结构中每条边上皆有概率分值。

在Probase知识库的基础上，申请人的研究小组和微软合作开发了以下几个基于文本理解和分析的应用：

* 万维网上top-k类型列表的抓取和分析（发表在KDD 2012、ICDE 2013，已申请中国发明专利）
* 基于上下文的万维网图片搜索和聚类（发表在EDBT 2013、ECML 2014）
* 词语相似度计算（发表在CIKM 2013)
* 基于概念（话题）的Web语义搜索工具（发表在ER 2012）
* 针对Web上表格的搜索引擎（发表在ER 2012）
  1. **文本中比喻的识别**

我们识别文本中的隐喻并通过找出比喻中的本题和语体来解释其中的比喻。通过句式匹配，例如：，和约束规则来提取文本中可能的本体、喻体关系对，并利用了一个通用的、概率性的IsA知识库过滤其中的非比喻成分、扩充其中的隐喻成分，构建了一个概率性的隐喻数据集。

对于一个未知的IsA语句，根据词性、语句结构等特征，提取中其中可能作为本体和（或）喻体的成分。利用IsA数据集对本（喻）体进行扩充和解释，依据隐喻数据集判断是否为含有隐喻的句子并对其中的本体和预提作出解释。该项工作第一次提出了隐喻解释的问题，并采用了大数据驱动的非监督式学习方法，发表在TACL Vol. 1 2013上。

* 1. **二元关系模式推理**

我们把开放式信息提取系统例如reverb中的三元组中的主语和宾语链接到知识图谱例如Freebase中的实体上。对于每一个reverb相近意义的谓语，都可以得到一个主语集合和宾语集合。因为诸如freebase的知识图谱往往有一套类型的层次结构，每一个实体可以对应多个不同层次上的类型，所以我们能通过把主语宾语的链接得到一个谓语两端相应类型组合（称之为模式）的概率分布。基于使用尽可能具体的模式匹配尽可能多的三元组的原则，我们就能针对一个自然语言的谓语，得到它两端最有可能出现的主宾语类型。最后的结果比传统的selectional preference有明显提高，该项工作发表在EMNLP2015上。

* 1. **其他文本理解、处理研究**

此外，我们的研究小组还在命名实体链接（又称为Wikification, CIKM 2013）和新浪微博谣言识别等实用性很强的文本理解应用方面有所突破，大量利用Wikipedia以及社交媒体中的图结构作为高维特征，设计了一系列与图数据相关的算法，有效解决实际问题。这些算法和经验都值得本项目借鉴。

2．**工作条件**（包括已具备的实验条件，尚缺少的实验条件和拟解决的途径，包括利用国家实验室、国家重点实验室和部门重点实验室等研究基地的计划与落实情况）；

本项目的依托单位为上海交通大学计算机系。它拥有国家级一级学科“计算机科学与技术”、五位千人计划获得者、五位国家杰出青年基金获得者，并与微软共建教育部级重点实验室， 拥有良好的科研氛围与研究条件。特别是近年来，该系通过引进大批海外留学归国专家，实力得到进一步加强，在2014 年发布的Academic Rankings of World Universities 继续保持全球前100 名。教育部－微软智能计算重点实验室是交通大学和微软亚洲研究院在多年良好合作的基础上为实现使未来的计算机和机器人能够看、听、学，能以自然语言的方式与人类交流这一共同使命而成立的。实验室成立后，学校和微软亚洲研究院在人力资源、实验室场地、研究经费等诸多方面给予了大力支持。截至2014 年，联合实验室共发表论文400 余篇，其中以交大与微软共同研究成果为基础的论文30余篇；双方教授与资深研究员互访与学术交流20 余人次。目前实验室拥有交通大学的高级研究人员26 名，微软亚洲研究员的研究人员16 名，博士研究生80 余名。

本申请人领导的数据与程序语言技术(ADAPT)实验室，作为教育部—微软重点实验室的组成部分，将获得重点实验室在软硬件资源、工作场所和硕士生、博士生人才方面的全面支持。实验室现有博士生5人，硕士生12人，在信息抽取、数据处理和知识构建等方面取得重要成果。拥有高性能服务器集群等先进的硬件平台。实验室和微软亚洲研究院、新加坡国立大学、POSTECH、香港理工大学、普林斯顿大学以及美国 Google公司有紧密和广泛的合作关系。不仅如此，上海交通大学计算机系专门为本申请人提供实验室场地和工作设施。在项目进行中，我们将购置数台大容量数据服务器共实验使用。我们也会为团队成员添置各种IT设备 。

3．**承担科研项目情况**（申请人和项目组主要参与者正在承担的科研项目情况，包括国家自然科学基金的项目，要注明项目的名称和编号、经费来源、起止年月、与本项目的关系及负责的内容等）；

**朱其立承担的科研项目**

* 阿斯利康合作研究基金，《医药文本挖掘》，2013-2015，负责人。医药文本，如电子病历中存在大量因果关系，例如症状的诊断、服药后的作用等。本项目的研究成果可以被应用到医药文本分析和理解上。
* 自然科学基金面上项目，批准号61373031，《基于动作概念的本体知识库及在文本处理上的应用》，2014年－2017年，负责人。为本项目提供与文本中事件抽取相关的知识与数据支持。
* 自然科学基金中韩合作交流基金，国科金外资助字第(61411140247) 号，《多语言、跨文化语义联想概念网络的研究》2014年－2015年，负责人。 为本项目提供图算法基础。另外联想网络中的关系和概率可以被用作因果推理中的一个特征。

**Dominik Scheder承担的科研项目**

无

**Seung-won Hwang承担的科研项目**

* 自然科学基金中韩合作交流基金，国科金外资助字第(61411140247) 号，《多语言、跨文化语义联想概念网络的研究》2014年－2015年，参与人。

4．**完成国家自然科学基金项目情况**（对申请人负责的前一个已结题科学基金项目（项目名称及批准号）完成情况、后续研究进展及与本申请项目的关系加以详细说明。另附该已结题项目研究工作总结摘要（限500字）和相关成果的详细目录）。

申请人已结题的青年基金项目《自动推断海量定制数据格式研究》（批准号61100050）已于2014年12月圆满完成。在这个项目中，我们不仅完成对半结构化文本数据的格式自动生成及内容抽取工作，还将研究衍生到非结构化数据的信息抽取与从属关系知识库Probase的构建上。该项目发表了十多篇高水平学术论文，申请并授权了多项国家专利，同时获得两项国际上颇有竞争力的奖项。从属关系的抽取只是一个开始，这个项目的后续研究包括了动作概念抽取、自由联想网络抽取以及本次申请的因果关系网络抽取等一系列知识获取和知识库构造研究工作。这些工作试图从不同侧面解决自然语言文本理解中的难题。青年基金项目开启了申请人对文本大数据处理与理解的研究热情，项目执行中积累的文本分析与自然语言处理的经验、数据、算法和源代码都将使本次申请的常识性因果推理项目获益。

**项目结题报告摘要**

本项目在对文本半结构化数据做增量格式分析的研究基础上，探求了对非结构化文本的信息抽取以及知识发现方面的工作。主要工作包括以下几个方面：1)递增式定制数据格式抽取；2）自动抽取IsA关系并构建 Probase知识库；3）利用Probase知识库自动理解网页表格、top-k列表、文本中比喻、计算词语间相似度等自然语言处理问题；4）命名实体链接；5）基于概念的网页搜索及图像搜索算法；6）数据隐私保护算法。一共发表高水平国际会议与期刊论文14篇，申请中国国家专利7项，获得Google Faculty Research Award一项，著名数据库会议DASFAA 2014最佳论文奖一项。在该项目的支持下，培养本科生15名，硕士毕业生6名，博士在读生2名。多名学生继续出国深造或被跨国公司录用。总的来说，项目成果丰硕。

**相关成果目录**

1、 Kaiqi Zhao、Zhiyuan Cai、Qingyu Sui、Enxun Wei、\*Kenny Q. Zhu ，Clustering Image Search Results by EntityDisambiguation ，ECML 2014, 2014/10/10-2014/10/12, 2014/10/10, 口头报告 ，EI ，已标注会议论文

2、 Xiao Jia、Chao Pan、Xinhui Xu、\*Kenny Q. Zhu、Eric Lo ，Rho-uncertainty Anonymization by Partial

Suppression. ，19th International Conference on Database Systems for Advanced Applications (DASFAA 2014), 2014/4/1-2014/4/3, 2014/4/1 ，EI ，已标注 会议论文

3、 Jack Wei Sun、Franky、\*Kenny Q. Zhu、Haixun Wang ，Query Suggestion by Concept Instantiation. ，ISWC 2013, 2013/10/21-2013/10/25, 2013/10/01 ，EI,ISTP ，已标注 会议论文

4、 Peipei Li、Haixun Wang、\*Kenny Q. Zhu、Xindong Wu、Zhongyuan Wang ，Computing Term Similarity by Large Probabilistic isA Knowledge. ，CIKM, 2013/10/31-2013/11/2, 2013/10/01 ，EI,ISTP ，已标注 会议论文

5、 Zhiyuan Cai、Kaiqi Zhao、\*Kenny Q. Zhu、Haixun Wang ，Wikification via Link Co-occurrenc，CIKM, 2013/10/31-2013/11/02, 2013/10/01 ，EI,ISTP ，已标注 会议论文

6、 Hongsong Li、Kenny Q. Zhu、Haixun Wang ，Data-Driven Metaphor Recognition and Explanation. ，Transactions of ACL, 1卷, pp 379-390, 2013/10/01 ，EI,SCI ，已标注 期刊论文

7、 Kai Sun、\*Kenny Q. Zhu、Yan Huang、Xiaobin Ma ，Watermarking Road Maps against Crop and Merge Attacks. ，ACM IH&amp;MMsec, 2013/06/17-2013/06/19, 2013/06/01 ，EI,ISTP ，已标注 会议论文

8、 Zhixian Zhang、\*Kenny Q. Zhu、Hongsong Li、Haixun Wang ，Automatic Extraction of Top-k Lists from the Web，IEEE 29th International Conference on Data Engineering, ICDE 2013, 2013/04/08-2013/04/11, 2013/04/01，EI,ISTP ，已标注 会议论文

9、 Kaiqi Zhao、Enxun Wei、QIngyu Sui、\*Kenny Q. Zhu、Eric Lo ，CISC: Clustered Image Search by Conceptualization. ，16th International Conference on Extending Database Technology, EDBT2013, 2013/03/18-2013/03/21, 2013/03/01 ，EI ，已标注  会议论文

10、 Jingjing Wang、Haixun Wang、Zhongyuan Wang、\*Kenny Q. Zhu ，Understanding Tables on the Web. ，ER,10/10/2012-12/10/2012, 2012/10/01 ，EI ，已标注  会议论文

11、 Yue Wang、Hongsong Li、Haixun Wang、\*Kenny Q. Zhu ，Concept-Based Web Search ，ER, 10/10/2012-12/10/2012,2012/10/01 ，EI ，已标注  会议论文

12、 Zhixian Zhang、\*Kenny Q. Zhu、Haixun Wang ，A System for Extracting Top-K List from the Web, KDD, 2012, EI ，已标注  会议论文

13、 \*Kenny Q. Zhu、Kathleen Fisher、David Walker ，LearnPADS++ : Incremental Inference of Ad Hoc Data Formats. ，PADL, 2012, 2012 ，已标注  会议论文

14、 Wentao Wu、Hongsong Li、Haixun Wang、Kenny Q. Zhu ，Probase: A Probabilistic Taxonomy for Text Understanding. ，SIGMOD, 2012, 2012 ，EI,SCI ，已标注  会议论文

**（三）资金预算说明**（购置单项经费5万元以上固定资产及设备等，须逐项说明与项目研究的直接相关性及必要性。）

无单项超过5万元的固定资产。

**（四）其他需要说明的问题**

无。