分别读了 Knowledge fusion、Knowledge Vault、Freebase 三篇。其中 from data fusion to knowledge fusion 研究了应用数据融合技术解决知识融合的过程，选择了三种已有的 data fusion 技术，并进行了调整改进使得更好的应用于知识融合。Knowledge Vault 则是基于 knowledge fusion 这篇 paper 中涉及的一些技术来构建了 web 规模的概率知识库。这两篇中均涉及到了使用已有的知识库中的先验知识,而使用的已有知识库就是具有大量手工验证三元组的 freebase。读的第三篇论文就是 Freebase 的一个 demonstration。

Knowledge fusion 这篇 paper 属于将已有技术进行调整用于解决新的任务，其中对现有技术的多种方法的进行选择时，首先明确了任务目标（需要给三元组一个真实性概率），于是分析利弊选择了易于得到可靠概率的方法。除此之外为了便于应用，将三维的知识融合任务合并（extractor，url）降维与数据融合维度一致的二维。我认为都是实践中非常值得学习的办法。

Knowledge fusion 侧重技术的演变过程，Knowledge Vault 这篇更侧重介绍如何利用自动构建 web 规模的概率知识库，并通过已有的知识库中的先验知识训练先验模型，对候选三元组评估一个置信度，来提高从 web 源获取知识的质量。基于图的先验计算使用了两种方法，PRA 和 MLP，前者训练了两种二元分类器。后者将链接预测问题看作矩阵。在 discussion 中提到了几点想法，如何建模事实的互斥性和软相关性，一个人不可能有两个出生地，一个人的出生日期通常会比孩子早 15-50 年。以及在论文中可以发现，未对 freebase 中不存在的实体进行处理，这一类实体如何添加并处理与之相关的关系？

论文一：From data fusion to knowledge fusion

1. 文章解决了什么问题：数据融合是从不同可靠性的不同来源中提取的多个观测值中，识别数据项的真实值；知识融合则是识别从多个信息源抽取的主谓宾三元组真实值。这篇文章将达到 SOTA 的数据融合技术应用于知识三元组的 knowledge base（比之前的 papers 中使用的数据集大三个数量级的融合文件），展示数据融合方法解决知识融合问题的巨大前景，并通过对方法的误差分析提出研究方向。研究如何使现有数据融合技术适应自动构建大规模

knowledge base。

2. 为什么要研究这个问题：从多个可能冲突的数据源中抽取信息并协调，确保真实值存到数据仓库中非常重要。
3. 文章的主要贡献：（1）定义知识融合，采用现有的三种数据融合技术解决知识融合，并设计基于 MapReduce 的高效实现，在数据是之前 paper 中数据量的 1000 倍的知识库中评估性能。（2）对现有方法进行简单的改进，显著的提高了质量，尤其是估计概率的校准质量（即确实能达到预测概率与实际准确度很大程度上一致）。（3）对本文的方法进行了错误分析，并列出建议的下一步研究方向。
4. 相关工作：（1）考虑到了相比数据融合，知识融合中更易出现的新增 noise。即：知识融合可视化比数据融合可视化要多考虑一维 extractors。（如下图）extractors 将 raw data 转换为结构化知识的过程主要有三步：识别数据中哪部分是数据项和对应的值；将涉及到的实体链接到对应的实体标识符；将数据中涉及到的关系链接到对应的知识库 schema 中。这三步中易导致的错误与数据融合不同，增加了噪声。（2）与前人的工作不同，本文知识库构建着重使用无监督方法，将 extractors 视为黑盒。（3）本文从大量领域中各种类型的数据中抽取的知识来检测错误，而不是以领域为中心。
5. 数据融合的方法：早期是基于规则，如选择最新更新的源，或计算平均值/最大值/最小值，这些方法可以通过使用数据库查询来提高效率。最近的基于无监督学习或半监督学习的高级解决方案分为三类：

   （1）基于 voting：选择一个由最大量的源提供的值。（2）基于质量：选择可信度高的。（根据如何度量可信度分为四种办法：基于 web 链接、基于 IR、贝叶斯方法、图形模型方法）（3）基于关系：对基于质量的扩展，多了要考虑源之间的关系。
6. 知识抽取系统：

   使用预定义的 Freebase 类别，每个为此也是 freebase 中的预定义谓词。三元组（主语，谓语，宾语），（主语，谓语）对应数据融合中的“数据项”，宾语视为数据项的值。相比于数据融合保留来源，知识融合保留更丰富的信息，包括使用的哪个抽取器抽取到的，抽取的上下文是什么，抽取器每次抽取的

置信度等。

数据来源：文本文档（txt）、DOM 树、Web 表（TBL）、Web 注释（ANO）

任务：三元组识别-实体链接-谓词链接

extractors 对不同类型的数据来源应用了不同的技术。

TXT：（1）使用标准的自然语言处理工具来进行命名实体的识别、解析等

（2）使用 freebase 的三元组作为训练数据进行远程监督

DOM 树：用相似的方法远程监督，此外从 DOM 树的结构中获取特征。

TBL:采用达到 SOTA 的 schema 映射技术将表的列映射到 freebase 谓词。

ANO：从 schema.org 中的本体到 freebase 中的本体的半自动定义的映射。

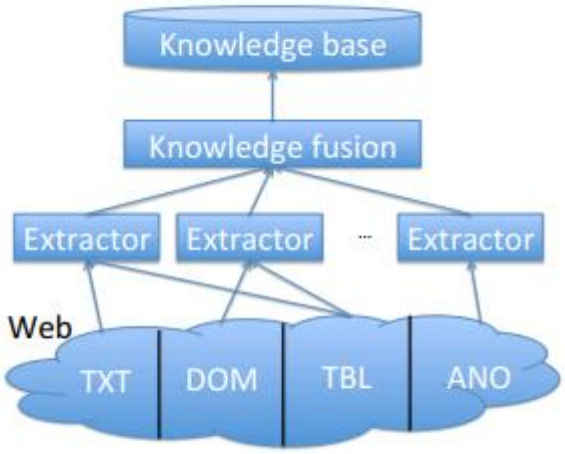

Figure 2: Architecture of knowledge extraction and fusion.

不同的 extractor 可以不同的顺序执行三个基本任务，也可以合并执行。Extractor 可以相互关联，采用相同的基础技术，使用相同的实体链接工具，但发挥不同的功能。

7. 知识抽取的质量评估：如果 freebase 中出现（s，p，o），则正确；未出现（s，p，o）但出现了（s，p）则错误。由于大多数谓词只有 1 或 2 个真实值，所以基于 LCWA 假设的评估效果很好。
8. 知识抽取的三个挑战：

（1） 三元组是 extractor 从数据源中抽取的，错误不仅仅是来自数据源本

身，在三元组识别、实体链接、谓词链接中都有可能出现错误。

（2） 数据融合是对数据项做出决定，非 0 即 1，非对即错。而知识融合的输出的三元组的真实性概率。

（3） 知识融合需要处理比数据融合大几个数量级的数据。

9. 如何将数据融合技术应用到知识融合：

方法选择：由于知识融合的目的是计算每个三元组的真实性概率，所以选择了易于得到有意义概率的方法：VOTE、ACCU、POPACCU。

Adaptions:(1)首先 DF 的输入是二维矩阵,KF 是三维,多了一维 extractor，将（extractor，URL）看作一个数据源，减小 KF 输入的维度。(2) 其次，DF 的输出是对每个来源提供的值的决策，而 KF 的输出是对三元组的真实性概率，ACCU、POPACCU 选择贝叶斯分析计算，VOTE 则是用（s，p，o）的个数 m/（s，p）的个数 n。(3) 基于 MapReduce 框架的知识融合体系结构如图：

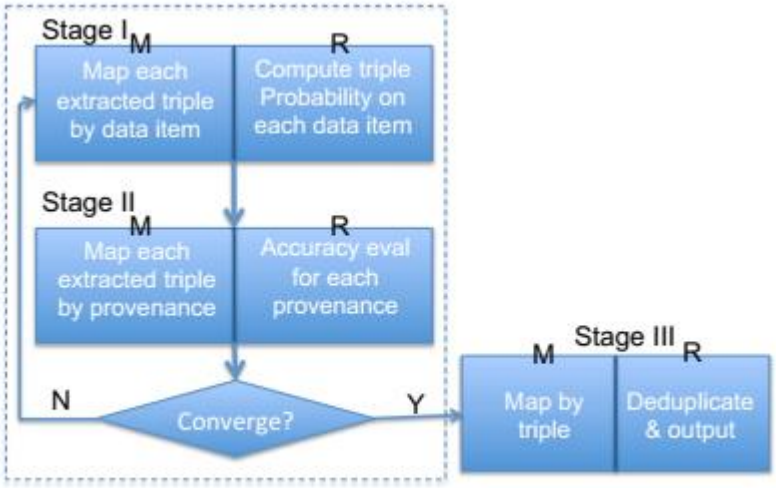

Figure 8: MapReduce implementation of ACCU and POPACCU.

三个阶段，每个阶段包括 Map 和 Reduce 两步，ACCU 和 POPACCU 需要迭代前两个阶段，直到收敛为止。VOTE 不需要迭代，仅具有阶段 1、3。

使用采样+R 轮（默认为 5）强制终止的方法加速执行。

10. 三种方法的评估：

VOTE 存在的问题：预测概率通常低于真实的概率、数据项只有 1 或 2 个出处时 vote 不足。

ACCU 和 POPACCU：对概率较高的三元组高估，对概率较低的三元组低估。

改进办法：(1) 适当的粒度选择，site-level、predicate-level、pattern-

level（2）忽略低质量数据来源（3）利用高质量知识库预测新抽取的三元组

11. 未来方向：（1）区分由 extractor 导致的错误和由 web 源导致的错误，可独立的评估二者质量，确定潜在错误以避免发生。（2）识别 web 源之间、extractor 之间复杂的关联。（3）非功能性谓词的处理。（4）需要一种可以推理层次和实体相似度的策略，而不是仅集中在字符串等相似性上。（5）利用抽取结果的置信度过滤三元组。（6）直接过滤掉样本量小的来源可能导致三元组没有任何概率估计，要合理的利用低覆盖率的数据来源。（7）改进封闭世界假设。（8）开放领域的知识融合。

论文二：Knowledge Vault

1. 文章简介：为了进一步扩大知识库规模，需要探索自动化构建知识库的方法。文章介绍了一种应用了自动构建 web 规模的概率知识库的新方法的 Knowledge Vault。
2. Knowledge Vault：从 web 内容（TXT、TBL、DOM、ANO）中抽取的内容与现有知识库中的先验知识相结合，使用有监督的机器学习方法来融合不同信息源。使得 Knowledge Vault 规模大，且具有概率推断系统，该系统可以计算事实正确性的校准概率。即每个三元组有一个相关联的置信度，代表 KV 认为该三元组是正确的概率。
3. 构建知识库方法的发展：旧方法是人工+结构化知识库集成。这样会导致只有常见的实体的常见属性（head content），而且人工的知识有限。需要采用新方法，自动从 web 中抽取，但直接使用该方法会产生许多错误，所以本文利用知识库（Freebase）中已有的知识建立事实正确性的先验模型。
4. 贡献：（1）KV 将从 web 上抽取到的有噪声的数据与 freebase 中获得的先验知识相结合，先验模型可以帮助克服抽取过程中的错误和数据源本身的错误。（2）KV 比其他知识库规模大很多。从各种 web 源（TXT、DOM 树、TBL、ANO）中抽取了事实。（3）对不同抽取方法的质量和覆盖率进行比较，评估封闭世界假设的有效性。
5. KV 的主要组成部分：extractors、基于图的先验概率计算系统、知识融合系

统。

Extractors 从 web 源抽取三元组，每个 extractor 为抽取到的三元组分配置信分数。基于图的先验概率计算系统基于 freebase 中已经存在的三元组学习每个候选三元组的先验概率。知识融合系统根据不同 extractor 与先验计算器之间的协议，计算三元组为真的概率。

6. 分类器训练集和测试集的标签如何确定：基于局部封闭世界假设（LCWA），上篇文章有总结，一个给定候选三元组，如果 freebase 中出现（s，p，o），则正确；未出现（s，p，o）但出现了（s，p）则错误。未出现（s，p）则从训练集测试集中丢掉。
7. 抽取方法：

   TXT：使用标准的 NLP 工具执行 ner、词类标注、依赖分析、实体链接等，再用远程监督训练关系抽取器。

   训练关系抽取器：对每个谓词，从 freebase 中抽取具有该谓词的实体对，找到抽实体对的原句子，总结句子的 features 和 patterns。根据 LCWA 标注的数据作为训练集，拟合二分类器。

   DOM：与 TXT 情况一样训练分类器，除了 features 是从 DOM 树中而不是文本中获得。

   TBL&ANO：与上篇文章一致，不再总结。
8. 基于图的先验模型（可看做是预测图中链接存在的可能性）采用两种方法再进行融合。

   （1） path ranking algorithm

   统计谓词 p 链接的实体对，路径能走通的实体对认为是成功的，将路径作为规则，一对实体对之间可能有多个路径（规则），拟合一个二元分类器来合并得到公共路径。再为每个谓词以不同路径为 features 拟合一个二元分类器。

   （2） 神经网络模型——使用标准的多层感知器（MLP）。
9. Extractors 融合/extractors 和 priors 融合：为抽取的三元组构造特征向量，特征向量为 extractor 抽取（该三元组的来源数目的平方根，抽取三元组用到的所有源数据抽取器打的分的平均分）。为每个谓词单独设置分类器，

可以根据不同的可靠性建模。训练融合系统。当源数据越多，真三元组的先验概率约趋近于 1，假三元组的先验概率越趋近与 0.5，但是一直低于 0.5。

10. 自动构建知识库文献分类：（1）基于 wiki infobox 和其他结构化数据源的构建（2）对整个网络使用无 schema 的开放信息抽取技术。（3）对整个网络使用固定的 ontology/schema 进行抽取。（4）构造了 is-a 的 taxonomies 的知识库。
11. 图的链接预测方面文献：（1）使用离散的马尔可夫随机字段或直接对变量之间相关性进行建模的方法（2）使用潜在变量通过离散因子或连续因子间接建模相关性的方法（3）使用算法方法近似相关的方法

论文三：Freebase

Freebase 这篇内容比较少，我在网上找了一些资料如下：

参考地址：https://developer.aliyun.com/article/717320

1. Freebase 的结构分为三层：Domain -> Type -> Topic。

   1）在 Freebase 中，每个条目叫做一个 Topic，每个 Topic 中的固定字段，叫做“属性”（Property）；

   2）所有同类的 Topic 组成一个 Type，比如所有电影 Topic 就属于同一个 Type，每个 Type 都有一套固定的 Property，因此同类信息可以直接比较和关联；

   3）所有相关的 Type 组成一个“域”（Domain）

2. 整个 Freebase 数据存储是一张大图，每个点都使用 type/object 定义，边使用 type/link 定义，不管是模型还是 Topic，其数据都作为点存储于图数据库中，通过边进行关联。

Freebase 的后台数据库 Graphd 以点和点间关系（边）的图状结构来组织数据，通过二进制数据存储来储存点和边，并以哈希表的方式存储组织数据，它在用户下载上传数据时的起到临时数据缓存器的作用，对数据进行检验处理后，再保存到 Graphd 中。

数据库中以数组的方式对点及其关系的元数据进行建模，以表格的形式进行存储，表格中的每条数据对应一个点边数组。点边数组一般由 4 个主要的数组成员组成，分别是源点、关系、目标点、源点值，于是点边表中就按这四个成员设定相应的四列。

Graphd 的图是有向图，边的方向从源点指向目标点，但执行查询时，可向前或向后遍历所有边来获取结果。

3. Freebase 主要是从维基抽取结构化数据并发布成 RDF，是完全结构化的，但数据来源不局限维基，还导入了数量众多的专业数据集，并提供数据查询和录入机制。DBpedia 与 Freebase 类似，但 DBpedia 存在数据结构化程度低、数据不一致、数据质量低、数据权威性和规范性不高、对 Wikipedia 数据的更新存在滞后性等问题。两者的数据在 08 年 11 月通过内置 OWL 属性 owl:sameAs 进行互联，使得双方数据集关联更密切。