## Freebase: A Collaboratively Created Graph Database For Structuring Human Knowledge

这篇文章发表于SIGMOD2008，主要介绍了Freebase的一些特点，如下：

* 存储在一个可伸缩元组中
* 一个基于HTTP/JSON的API
* 一个轻量的、协作的类型系统
* 一个庞大的、多样化的数据集
* 显式规范化的本体

因为这篇文章只介绍了Freebase数据集的特点，但是没有交代构建Freebase的具体细节和本体结构，所以搜集了一些关于Freebase的相关信息。

Freebase 是一个由元数据组成的大型合作知识库，内容主要来自其社区成员的贡献。它整合了许多网上的资源，包括部分私人wiki站点中的内容。Freebase 致力于打造一个允许全球所有人（和机器）快捷访问的资源库，由美国软件公司Metaweb开发并于2007年3月公开运营。2010年7月16日被Google收购， 2014年12月16日，Google宣布将在六个月后关闭 Freebase ，并将全部数据迁移至Wikidata。

#### 基础模型

* Topic：即实例或实体，每一条信息叫做Topic，比如：姚明等。
* Type：类型或概念，每个Topic可以属于多个Type，比如：人、运动员等。
* Domain：域，对类型的分组，便于schema管理，比如：人物。
* Property：属性，每个Type可以设置多个属性，其值默认可以有多个，可通过设置unique为true限制只能有一个值。比如：出生日期、所在球队等。属性值类型可以是基本类型，比如：整型、文本等；也可以是另一个type，比如：所在球队、父母等，这种情况叫做CVT，compound value type 组合值类型，比如：所在球队就是一个CVT，它有自身结构化的属性，不仅仅只是一种简单的值。
* MID：实体编号。不考虑实体合并和分裂时，一个实体和一个MID是一一对应的；当考虑实体合并和分裂时，多个MID可能指代一个实体，但是只有一个MID是master，其他的MID通过一个特殊的属性(<人.运动员.replaced\_by>)指向这个MID。
* KEY：可以通过key来唯一确定一个实体，一个实体可以有多个key，每个key都属于一个namespace，比如： "/en/yao\_ming"的namespace为"/en"、"/wikipedia/zh-cn\_title/姚明" 的namespace为"/wikipedia/zh-cn\_title"。对于平台基础模型的实体（Domain、Type、Property），Freebase会从Key中选一个值，作为该实体的ID。
* 属性约束：用于约束属性的取值范围，比如：类型约束（整型、文本、浮点型、datetime、CVT等）、条件约束（是否单值、是否去重、主属性、逆属性等）。

#### **MQL**

Freebase的主要查询语言为Metaweb Query Language（MQL），一种易于使用的面向对象的查询语言，具有基于树的对象结果结构。我参考了Freebase的官网说明，对Freebase进行了简单的查询，MQL用于提供HTTP API来进行Freebase的读写操作，其入参和反参都是json格式，使用起来较为方便

主要为以下几种实例：

1. 根据guid进行查询。

{"query": [

{

"limit": 1,

"guid": "42caf4fa-4725-11e5-8fdc-f80f41fb03aa",

"name": "", //object的字段名

"/tv/tv\_actor/date:"" //类型tv\_actor的属性出道时间date

"/tv/tv\_actor/tv\_program": [

{

"name": "",

"limit": 3

}

]

}

]}

1. 根据类型的guid或ID查类型schema

{"query": [

{

"limit": 1,

"guid": "42caf4fa-4725-11e5-8fdc-f80f41fb03aa",

"type": "/tv/tv\_actor"

"/type/domain/types": ""

}

]}

1. 根据类型ID的实例列表，并指定需要字段列

{"query": [

{

"type": "/film/film",

"guid": "", //需返回guid字段值

"name": "" //需返回name字段值

"limit": 20, //指定返回条数，若返回所有则不指定该字段

}

]}

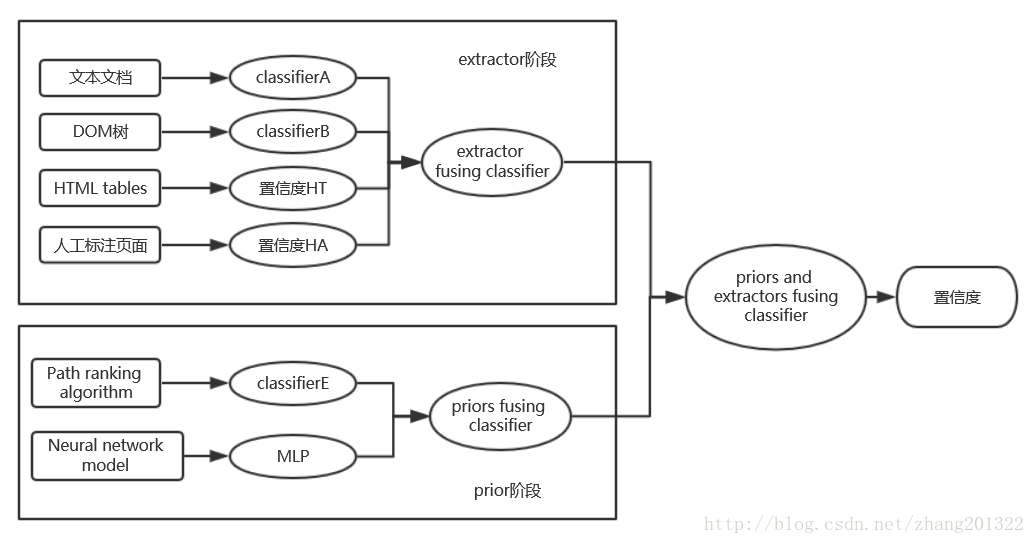
## Knowledge Vault: A Web-Scale Approach to

## Probabilistic Knowledge Fusion

这篇文章是Google公司发表在KDD2014的一篇论文，截止到目前引用量高达1077，这篇论文主要是介绍如何构建一个完整的知识库，主要讲述了作者以FreeBase为基础自动构建一个互联网规模的概率知识库的过程。

基于机器学习，Knowledge Vault不仅能够从多个来源（文本，表格数据，页面结构，人工注释）中提取数据，而且还可以根据所有可用数据推断事实和关系。网络当然包含大量的错误数据，因此框架依赖于现有的知识库（例如Freebase），以便在评估过程的其中一个步骤中验证事实。研究人员将该过程描述为“图中的链接预测”，并试图通过采用两种不同的方法来解决它：a）路径排序算法（PRA）和b）神经网络模型（MLP）。

本文的整体架构主要分为提取器、先验模型和融合器这三大部分，按照论文的叙述总结如下图：



以下是这篇论文的一些创新点和思考点：

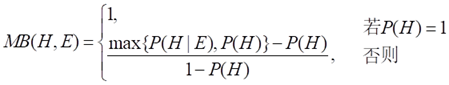
* 提出了概率三元组的想法

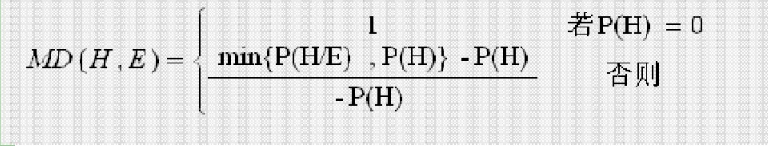
概率三元组是在常见的知识图谱三元组中加入一个概率值表示此三元组成立的概率，表示为Pr(G(s,p,o)=1|·)。概率三元组更为常见的说法叫做不确定性推理，是指建立在不确定性知识和证据基础上的推理。例如，不完备、不精确知识的推理，模糊知识的推理等。实质上是一种从不确定的初试证据出发，通过运用不确定性知识，最终推出具有一定程度不确定性但却又是合理或基本合理的结论的思维过程，与知识图谱中概率三元组意义相近。

概率三元组的不确定性可以分为两类：知识源的不确定性和推理过程的不确定性。

在不确定推理中有一个很重要的概念叫做知识强度，可以引入知识图谱中进行知识推理的约束，我们用CF表示知识的静态强度，CF(H, E)的取值为[-1, 1]，表示当E为真时，证据对H的支持程度，其值越大，支持程度越大。

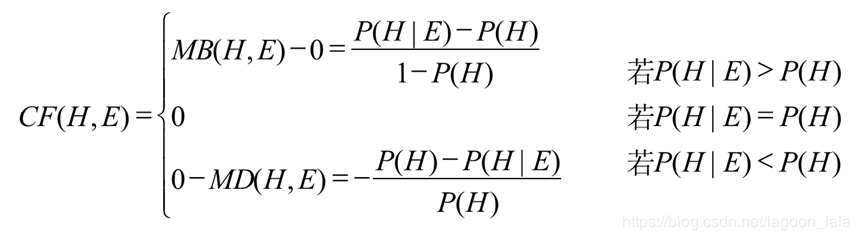
在CF模型中，把CF(H, E)定义为CF(H, E)=MB(H, E)-MD(H, E)，式中MB称为信任增长度，MB(H, E)和MD(H, E)的定义为





当MB(H, E)>0时，有P(H|E)>P(H)，即E的出现增加了H的概率；当MD(H, E)>0时，有P(H|E)<P(H) ，即E的出现降低了H的概率。

根据前面对CF(H, E)可信度 、MB(H, E)信任增长度、MD(H, E)不信增长度的定义，可得到CF(H, E)的计算公式：



CF(H,E)的优点是可以表达出证据无法支持推理或者对结论有反作用的情况，当MD大于MB时，证据会降低推理的置信度。

* 处理嘈杂的数据源问题

因为此文中选取Freebase知识图谱作为三元组来源，所以假定Freebase中出现过的