# Domain Ontology的定义

本文档定义需要从页面中抽取的信息，而不是存储结构。信息提取的目标在于：如何用足够的信息使得所提取的对象能映射到结构化的数据库上。

每个ontology除了定义相应的类别、概念、属性和关系外，每个对应的对象下 还需要可以灵活编辑的抽取模板（支持可编辑）；

要考虑到各个不同网站 对同一个实体的描述可能是不同的；

框架上做可以任意添加 “领域”，程序框架实现即插即用，即马上能接入抽取流程。

所有的Entity Page都可能需要定期更新；

Entity page包括：电影、音乐、博文(?)、论坛帖子、博客、微博等

**新闻类（每个属性有对应的抽取模板）**

|  |  |  |
| --- | --- | --- |
| URL |  | 记录此页面的地址 |
| 发表时间 |  | 强制 |
|  | 抽取模板（比如wrapper、正则表达式） |  |
| 新闻来源 |  |  |
|  | 抽取模板（比如wrapper、正则表达式） |  |
| 类别 | 如：[新浪财经](http://finance.sina.com.cn/) > [国内财经](http://finance.sina.com.cn/china/) |  |
| 标题 |  | 强制 |
| 正文 |  | 强制 |
| 领域 | 电影\音乐\电视\其它 |  |
| 作者 |  | 可空 |
| **HBase存储方案 (由房磊提供编程接口)** | | |
| News | 表名 |  |
| Primary key | url |  |
| Timestamp | Long.max\_value-publish\_date | 时间最新的排在最前面 |
| Info:source | 新闻来源 | Info表示column family |
| Info:time | 发表时间 | Time表示column label |
| Info:cate | 类别路径 |  |
| Info:author | 作者 |  |
| Info:domain | 领域 |  |
| Content:title | 标题 |  |
| Content:body | 正文 |  |
| Content:keyword | 也许以后有用 |  |

**博文类（并没有对博文提供评论数据的关联）**

|  |  |  |
| --- | --- | --- |
|  |  | HBase表名blog |
|  | Long.max\_value-time | timestamp |
| 自定义分类 | 分类： [云南地理](http://blog.sina.com.cn/s/articlelist_1232204102_3_1.html) | Info:cate |
| URL | 记录此页面的地址 | Primary key |
| 标签 | [云南干旱](http://uni.sina.com.cn/c.php?t=blog&k=%D4%C6%C4%CF%B8%C9%BA%B5&ts=bpost&stype=tag)  [抗旱救灾](http://uni.sina.com.cn/c.php?t=blog&k=%BF%B9%BA%B5%BE%C8%D4%D6&ts=bpost&stype=tag) [肖育文](http://uni.sina.com.cn/c.php?t=blog&k=%D0%A4%D3%FD%CE%C4&ts=bpost&stype=tag) [杂谈](http://uni.sina.com.cn/c.php?t=blog&k=%D4%D3%CC%B8&ts=bpost&stype=tag) | Content:keyword |
| 标题 |  | Content:title |
| 发表日期 |  | Info:time |
| 正文 |  | Content:body |
| 作者（博主） |  | Info:author |
| 作者URL |  | Info:author\_url |
| 阅读/点击数 | （需定期更新） | Stat:nReadorClick |
| 评论数 | （需定期更新） | Stat:nComment |
| 收藏数 |  | Stat:nFavorit |
| 转载数 |  | Stat:nRepost |

**电影 （可参考各个档案库+douban+mtime做一个属性超集）**

|  |  |  |
| --- | --- | --- |
|  |  | Hbase表名：movie |
|  | (能直接是电影名吗？) | Primary Key: URL |
| 名称 |  | Info:name |
| 导演 |  | Info:director |
| 编剧 |  | Info:editor |
| 主演/演员 |  | Info:stars |
| 制片 | （可空） | Info:producer |
| 类型 |  | Info:cate |
| 片长 |  | Info:length |
| 时间/年代/上市时间 |  | Info:pubTime |
| 国家和地区 |  | Info:country |
| 剧情 |  | Info:desc |
| **1-5star** | （需定期更新） | Stat:ratingStarb |
| **期待数（想看）** | （需定期更新） | Stat:nExpect |
| **好评数（dig number）** | （需定期更新） | Stat:nPosi |
| **评论数** | （需定期更新） | Stat:nComment |
| **播放数 （看过）** | （需定期更新） | Stat:nPlay |
|  |  | Review:userid  Review:title Review:content; Review:time; Review:rating;  Review:location |

Review 和实体的关联是否很容易做到；要考虑各种不同页面的情况；特别是Next page情况

**音乐专辑/歌曲**

|  |  |  |
| --- | --- | --- |
|  |  | Hbase表名：music |
|  | Primary key | URL |
| 演唱者： |  | Info:singer |
| 别名： |  | Info:alias |
| 流派 |  | Info:style |
| 类型 | (单曲 or 专辑) | Info:type |
| 名字： |  | Info:Name |
| 发行时间： |  | Info：time |
| **1-5star** | （需定期更新） | Stat:1star stat:2star |
| **期待数（想看）** | （需定期更新） | Stat:nExpect |
| **好评数（dig number）** | （需定期更新） | Stat:nPosi |
| **评论数** | （需定期更新） | Stat:nComment |
| **播放数 （看过）** | （需定期更新） | Stat:nPlay |
|  |  | Review:userid  Review:title Review:content; Review:time; Review:rating;  Review:location |

Review 和实体的关联是否很容易做到；要考虑各种不同页面的情况；特别是Next page情况

**影星、歌星、博主**

|  |  |  |
| --- | --- | --- |
|  |  | Hbase表名：blogger |
| 博主URL |  | Primary key |
| 博客等级 |  | Stat:level |
| 博客积分 | （需定期更新） | Stat:credit |
| 博客访问数 | （需定期更新） | Stat:nRead |
| 名字 |  | Info:name |
| 百科资料 | （在博客中有百度百科资料） | Info:wiki |
|  |  | Hbase表名：twitter |
| 微博URL |  | Primary key |
| 微博粉丝数 |  | Stat:nFollowers |
| 微博关注数 |  | Stat:nFollowing |
| 微博数量 |  | Stat:nTweet |
| 微博描述 |  | Info:desc |
| 地区 |  | Info:location |
| 用户标签 |  | Info:label |

**评论**

|  |  |  |
| --- | --- | --- |
| 关联对象 | URL ---所评论entity page的页面url | **不单独存储(???)**  **或者单独存储一个，作为备份；后续review与entity之间的关联通过别的任务计算** |
| 作者 |  | author |
|  |  |  |
| 标题 |  |
| 地域 | （可能无） |
| 时间 |  |
| 正文 |  |
| 评级（分数/满分） | （可能无） |

**用户**

|  |  |  |
| --- | --- | --- |
|  |  | Hbase表名：blogger |
|  |  | Primary key: 网站主域名+IDname |
| IDName |  | Info:name |
| 年龄 |  | Info:age |
| 地域 |  | Info:location |
| 性别 |  | Info:gender |
|  |  |  |

**论坛话题（论坛类可能单做模板进行抽取）**

|  |  |  |
| --- | --- | --- |
|  |  | Hbase表名：post |
|  |  | Primary key: url |
|  | Long.max\_value-info:time | timestamp |
| 标题 |  | Content:title |
| 正文 |  | Content:body |
| 作者 |  | Info:user |
| 发文日期 |  | Info:time |
| 阅读/点击数 | （需定期更新） | Stat:nRead |
| 回复数 | （需定期更新） | Stat:nReply |
| ID | Add by史兴 | Stat:ID |
| 父节点ID | Add by史兴 | Stat:parentID |