

一、 介绍:

Entity Linking, 识别给定文本中出现的命名实体 (Named Entity), 并映射到特定的知识库中唯一的实体。包括命名实体识别、消歧等工作。

主要涉及内容包括:

- (1) 服务器 CGI 支持;
- (2) LDA 文本主题提取, 尝试进行实体消歧;
- (3) 搭建 Solr 索引;

最终结果以 Chrome 插件 (Chrome extension) 的形式展示, 用户可以在浏览网页时选择网页一段文本, 直接点击插件, 识别选中文本中出现的 entity。

二、 运行环境

数据集: DBpedia 2014

描述: Unknown action "a"Extended Abstracts, ttl 格式

网址: <http://oldwiki.dbpedia.org/Downloads2014>

前端: Chrome 55.0 (64-bit)

67 服务器 (CGI):

- Core: 2 * Intel(R) Xeon(R) CPU E5620 v3 @ 2.40GHz
- OS: Windows Server 2008 R2 Standard
- Mem: 24.0 GB

93 服务器 (Solr 索引):

- Core: 6 * Intel(R) Xeon(R) CPU E5-2609 v3 @ 1.90GHz
- OS: Ubuntu 4.0 SMP Tue x86_64 GNU/Linux
- Mem: 64GB

其它处理代码主要环境为 OS X El Capitan 10.11.6 下 Python 2.7 与 gcc46 4.6.4;
以及 Ubuntu 4.0 下 Python 3.0。若 Python 版本不兼容, 一般做少量改动即可。

三、 功能演示：（附带功能截图）

（1） Entity Linking

对于选中网页上的文本，筛选出其中的实体，并结合对应扩展信息一并展示在右上方（插件下方）：



（2） 筛选

对于已经展示出来的实体，可以在输入框内输入条件，进行进一步筛选：



(1) 当前选中本文段主题分布 T:

消歧: 出现 entity1 (A B), entity2 (B C)两个实体时, 因为存在重合单词 B, 我们需要在 entity1 和 entity2 中做选择, 这里借助 entity1 和 entity2 的实体的主题分布, 分别于当前选择文本主题分布 T 进行点积运算, 取较大者。

(2) 处理获得本文段中所有可能的实体候选集

如一段文本[A、B、C、D], 我们认为[A、AB、ABC、B、BC、BCD、C、CD、D]都属于实体候选集。1/2/3-gram 分别表示长度为 1、2、3 的邻近词组成的整体实体。(与一般 gram 含义略有不同)

4.2.3 实体、主题分布、附加信息

(1) 获取所有单个单词的主题分布, 以便计算选中文本段主题分布 T;

我们将出现在选中文本段中的所有的单词的主题分布进行累加。

(2) 获取所有候选集中, 在数据库索引中存在的真正实体, 同时返回其主题分布信息与一些其它介绍性质的附加信息。

4.2.4 查找

对于第 3 步中获得的实体, 我们需要检查冲突、进行消歧。

首先, 我们计算每个实体与当前文本段主题分布的点积, 我们认为是可信度:

实体名称:	实体 1	实体 2	实体 3	实体 4	...
可信度:	0.7	0.654	0.534	0.2331	...

之后, 按照可信度从高到低排序, 我们先选择可信度最高的实体, 然后对于由单词 X1、X2 组成的实体(X1_X2), 我们将候选集中所有包含 X1 和 X2 的实体删除, 之后再选择可信度最高的实体。不断循环直到候选实体集为空。

最终获得需要返回的结果实体集(包括扩展信息, 不再需要主题分布信息)。

4.2.5 实体、附加信息

将第 4 步中获取的最终实体集返回到服务器。

4.2.6 实体、附加信息

将第 4 步返回的最终实体集返回到前端, 利用 Chrome 插件进行展示。

五、 总结：

内容涉及 Chrome 插件制作、服务器 CGI 支持、LDA 模型提取文本主题特征以及借助 Solr 搭建索引，最终结合各方面完成 Entity Linking 的目标。包括实体链接、实体消歧、实体冲突等不同程度的解决。

作者：宋军帅

时间：2016-01-12