面向中文知识库 zhishi.me 的链接预测与元组分类任务

引论

链接预测是图数据挖掘中的一个重要问题，其通过已知的网络结构等信息，预测或估算尚未链接的两个节点存在链接的可能性。这种链接既包含了对未知链接（exist yet unkown links）的预测也包含了对未来链接（future links）的预测。链接预测通常应用于社交网络中，其可以作为准确分析社交网络结构的有力辅助工具，具体地，可以帮助分析数据缺失的网络，还可以用于分析演化网络，即预测未来链接。举例说明，由于在线社交网络的快速发展，链接预测可以基于当前的网络结构去预测哪些现在尚未结交的用户“应该是朋友”，并将此结果作为“朋友推荐”发送给用户，还可以用于已知部分节点类型的网络中预测未标签节点的类型。这些链接预测的思想也可以应用于图数据中关系和实体的预测。

Zhishi.me 是一个开放的中文链接图数据库，可以看做是一个复杂网络用图来表达。本任务基于图数据挖掘中链接预测的基本思想，通过计算尚未建立链接的两个 entity 发生链接的概率，实现图数据中 entity 与 entity 之间的链接关系预测。举例说明，已知下面两个三元组<entity A，配偶，entity B>与<entity B，儿子，entity C>，可以预测出<entity A，儿子，entity C>；或者已知三个三元组<entity A，生产，entity C>，<entity B ，生产，entity D>，<entity C，竞争，entity D>，可以预测出<entity A，竞争，entity B>。上述举例说明了对 entity 之间关系的预测，同时链接预测也可以实现对属性值的预测，例如，预测 entity A 的属性值是什么，在三元组中<entity A，出生于，？>表示 entity A 出生于某个地方，当然在其它的三元组中“？”也可能表示的是概念(concept)。

另一个相关的任务是三元组分类（triple classification），其表示的意义是对于预测的三元组判断对与错，即一个三元组是正确的，那么 triple classification 是正确的，否则是错误的。例如，三元组<姚明，出生地，上海>的 triple classification 是正确的，而三元组<姚明，性别，女>的 triple classification 是错误的。三元组分类的正确与否直接关系图数据库的质量，同时对于知识推理、问答系统有着非常

重要的影响。

任务描述：

数据库 zhishi.me 是一个开放的中文链接数据集，它是由大量的三元组组成，组成类型是<S,P,O>的形式，其中 S 和 O 表示的是 entity，P 表示的是 S 和 O 之间的关系（relation），其数据格式为：

实验中所有的数据都来自于数据库 zhishi.me

输入输出：

训练输入：用于模型训练的 triplet 集合（Train，内部格式为<S, P, O>），entity2id 集合（E，实体以及对应的 id），以及 relation2id 集合（R，关系名以及对应的 id）。

测试输入 1：用于模型 link prediction 性能测试的 corrupted triplet 集合（Test，内部格式为<S, P, ?>或<?, P, O>）。

测试输入 2：用于模型 triplet classification 性能测试的 triplet 集合（Validate，内部格式为<S, P, O>，部分为错误元组）。

说明：上述 Train、Test、validate 集合内的所有 entity 均包含于 E，所有 relation 均包含于 R。

训练输出：不限内容规范及格式。

测试输出 1：根据 corrupted triplet <S, P, ?>或<?, P, O>，利用训练所得的模型对该 triplet 进行复原，按照可信度从高到低的顺序列举出该 triplet 的 200 位 candidate entitiy list。按照 Test 内 corrupted triplet 的顺序每行打印一组 list，entity id 之间空格隔开。结果保存至 test_result.txt 文件。

测试输出 2：利用训练所得模型对 Validate 内的每一组 triplet 进行分类，按照 triplet 的顺序分别标记为 0（错误）或 1（正确）。数字之间空格隔开。结果保存至 validate_result.txt 文件。

评测方法：

Link Prediction 测试（输入 test_result.txt）：

(1) Mean Rank（MR）：

对比未经 corrupted 的 Test 集合，统计每组 candidate entity list 中 correct entity 的位次，计算出平均值作为评分依据。显然地，MR 的值越小，性能越好。

$$MR = \frac{1}{n}\sum_{i=1}^{n} R(i)$$

其中 n 为 Test 中元组总数，R（i）为第 i 组 candidate entity list 中 correct entity 的位次。

(2) Hit@10

统计所有 candidate entity list 中 correct entity 出现在前 10 位的 list 数在总数中所占的比例。Hit@10 越高，模型 Link Prediction 的准确性越好。

$$Hit@10 = \frac{\sum_{i=1}^{n} hit@10}{n} \times 100\%$$

其中，hit@10 当 correct entity 在 list 中的位次小于等于 10 时为 1，大于 10 时为 0。

(3) Hit@3

统计所有 candidate entity list 中 correct entity 出现在前 3 位的 list 数在总数中所占的比例。Hit@3 越高，模型 Link Prediction 的精确性越好。

$$Hit@3 = \frac{\sum_{i=1}^{n} hit@3}{n} \times 100\%$$

其中，hit@3 当 correct entity 在 list 中的位次小于等于 3 时为 1，大于 3 时为 0。

Triplet Classification（TC）测试（输入 validate_result.txt）：

根据已知数据，统计 validate_result.txt 文件中分类结果的正确率。

$$TC = \frac{\sum_{i=1}^{n} correct}{n} \times 100\%$$

其中，correct 在分类正确时为 1，错误时为 0。n 为 Validate 中 triplet 的总数。

最终得分的计算公式为

$$Score = 30\times(1-\frac{MR}{200})+30\times(Hit@10)+10\times(Hit@3)+30\times TC$$

任务提交指南

每一个参赛队需提交的材料如下：

1. 运行结果文件 validate_result.txt 和 test_result.txt
2. 代码及文档
3. 方法描述文档

4、参赛队员信息：每个参赛队设置一名队长，参赛队员信息主要包括参赛队队长和队员的姓名、职业、公司（学校）、联系电话、邮箱等，每个参赛队的成员不得超过 5 人。

以上三个文件需在任务提交截止日期前发送至邮箱：wutong8023@163.com，liuhefei_425317807@163.com

邮件的标题为：“LPTC+参赛队名称+学校/公司名称”。

代码及其文档需打包成一个文件（tar，zip，gzip，rar 等均可），用 code.xxx 命名，要求提交所有的程序源代码包含必要注释及相关的配置说明（包括软件版本，第三包说明等），确保程序能够正确运行，且所得结果与提交的运行结果文件相符。

方法描述文档用 LPTC.pdf 命名，包含算法描述及参数设置，需用 pdf 格式存储（LNCS 风格的 Springer 出版物格式），页数不超过 5 页。

联系方式

关于面向中文知识库 zhishi.me 的链接预测与元组分类的任何问题，请联系；

王昊奋（whfcarter@ecust.edu.cn）

漆桂林（gqi@seu.edu.cn）

地点（中国•上海市梅陇路 130 号

华东理工大学自然语言处理与大数据挖掘实验室）

邮政编码 200237

（南京市江宁区 东南大学路 2 号

东南大学知识科学与工程实验室）

邮政编码 211189

重要时间点

2016 年 5 月 10 日，发布知识库数据

2016 年 5 月 20 日，发布训练数据（标注评论数据）

2016 年 8 月 8 日，发布测试数据

2016 年 8 月 15 日，评测完成，停止接受结果