Tutorial:实体关系发现框架Limes

1. **软件安装**

   1.1 获取Limes

   ```
   git clone https://github.com/dice-group/LIMES
   ```

   1.2 编译源码

   进入 `limes-core` 目录编译：

   ```
   cd limes-core
   mvn clean install
   ```

   创建可运行Jar文件： `mvn clean package shade:shade -Dcheckstyle.skip=true -Dmaven.test.skip=true`

   Jar文件目录： `limes-core/target/limes-core-VERSION-SNAPSHOT.jar`

   1.3 运行Jar文件

   ```
   cd target
   java -jar limes-core-1.0.0-SNAPSHOT.jar  config.xml
   ```

   config.xml是自定义的配置文件。

2. **数据准备**

   Limes输入文件格式包括SPARQL端点，以及CSV, NT, TURTLE等格式的本地文件。 所有信息都需要用三元组的形式记录，NT格式文件如下：

   ```
   <http://zhishi.me/zhwiki/resource/民族> <http://www.w3.org/2000/01/rdf-schema#label>  "民族"@zh .
   <http://zhishi.me/zhwiki/resource/戏剧> <http://www.w3.org/2000/01/rdf-schema#label>  "戏剧"@zh .
   <http://zhishi.me/zhwiki/resource/电影> <http://www.w3.org/2000/01/rdf-schema#label>  "电影"@zh .
   ```

3. **配置文件**

   使用Limes进行实体关系融合的关键步骤是写好配置文件，包括数据源，融合算法，融合条件等信息。

   3.1 Prefixes

   将命令空间Namespace缩写为前缀标签Prefix Label，便于后文书写。例如:

   ```
   <PREFIX>
       <NAMESPACE>http://www.w3.org/1999/02/22-rdf-syntax-ns#</NAMESPACE>
       <LABEL>rdf</LABEL>
   </PREFIX>
   ```

   - 可根据需要配置多个Prefixes。

   3.2 数据源 Data Sources

   Data Sources包括Source 和 Target， 配置格式都一样。

   ```
   <SOURCE>
       <ID>mesh</ID>
       <ENDPOINT>http://mesh.bio2rdf.org/sparql</ENDPOINT>
       <VAR>?y</VAR>
       <PAGESIZE>5000</PAGESIZE>
       <RESTRICTION>?y rdf:type meshr:Concept</RESTRICTION>
       <PROPERTY>dc:title</PROPERTY>
       <TYPE>sparql</TYPE>
   </SOURCE>
   ```

- ID: 自定义数据源名称。
- ENDPOINT: 数据源地址，可以是SPARQL端点，也可以是本地文件(需要绝对路径)。
- VAR: 参与实体相似度计算的变量，这个变量在Metric表达式中会使用。
- PAGESIZE： SPARQL端点每次查询返回的最大Triple数量，本地文件设置为-1。
- RESTRICTION： 参与实体相似度计算的Triple限制条件。
- 可对数据进行预处理，例如： `lowercase(rdfs:label)` ，表示将 `rdfs:label` 的宾语全部改为小写字母。

3.3 度量表达式 Metric Expression

```
<METRIC>
    trigram(x.label, y.title) | 0.8
</METRIC>
```

- 使用Data Sources中配置的变量进行计算。
- 多个Mertic Expression可以使用MIN, MAX, ADD操作符结合使用，例如：

`MAX(trigrams(x.rdfs:label,y.dc:title)|0.3,euclidean(x.lat|long, y.latitude|longitude)|0.5).`

- MAX的第一个子表达式：x的rdfs:label与y的dc:title之间的Trigram相似度大于或等于0.3。
- MAX的第二个子表达式：x中的点(x.lat, x.long)与y中的点(y.latitude, y.longitude)之间的欧几里得距离大于或等于0.5。
- MAX操作符取两个子表达式中的最大值最为相似度。
- 目前所有操作符只支持两个Expression结合，但可以嵌套使用。
- 还可以使用Bool表达式AND, OR, DIFF对计算结果进行过滤, 例如：
  `AND(trigrams(x.rdfs:label,y.dc:title)|0.9, euclidean(x.lat|x.long, y.latitude|y.longitude)|0.7)`
  该表达式将返回两个子表达式融合结果的并集。
- METRIC 支持的原子表达式有：Cosine、ExactMatch、Jaccard、Jaro、JaroWinkler、Levenshtein、MongeElkan、Overlap、Qgrams、RatcliffObershelp、Soundex、Trigram，更多信息可参考链接。

3.4 机器学习

融合计算可以选择Metric Expression指定相似性度量表达式，也可以选择机器学习自动计算。故 <METRIC>和 <MLALGORITHM>标签二选一填写。

```
<MLALGORITHM>
    <NAME>wombat simple</NAME>
    <TYPE>supervised batch</TYPE>
    <TRAINING>trainingData.nt</TRAINING>
    <PARAMETER>
        <NAME>max execution time in minutes</NAME>
        <VALUE>60</VALUE>
    </PARAMETER>
</MLALGORITHM>
```

- NAME: 算法名, 支持womabt simple,wombat complete,eagle。
- TYPE: 训练方式，支持supervised batch，supervised active，unsupervised。
- TRAINING: 训练集文件地址，该文件只能包括以 `<http://www.w3.org/2002/07/owl#sameAs>` 作为谓语的三元组。
- PARAMETER：训练参数配置，可参考下表：

ML Algorithm	Supported types	Parameter	Default Value	Note
WOMBAT Simple	supervised batch, supervised active and unsupervised	max refinement tree size	2000	
		max iterations number	3	
		max iteration time in minutes	20	
		max execution time in minutes	600	
		max fitness threshold	1	Range 0 to 1
		minimum property coverage	0.4	Range 0 to 1
		property learning rate	0.9	Range 0 to 1
		overall penalty weight	0.5	Range 0 to 1
		children penalty weight	1	Range 0 to 1
		complexity penalty weight	1	Range 0 to 1
		verbose	false	
		atomic measures	jaccard, trigrams, cosine, qgrams	
		save mapping	true	
WOMBAT Complete	supervised batch, supervised active and unsupervised	Same as WOMBAT Simple		
EAGLE	supervised batch, supervised active and unsupervised	generations	10	Integer
		preserve_fittest	true	
		max_duration	60	[1,Inf)
		inquiry_size	10	[1,Inf)
		max_iterations	500	[1,Inf)
		max_quality	0.5	[0.0,1.0]
		termination_criteria	iteration	enum
		termination_criteria_value	0.0	[0.0,Inf)
		beta	1.0	[0.0,1.0]
		population	20	[1,Inf)
		mutation_rate	0.4	[0.0,1.0]
		reproduction_rate	0.4	[0.0,1.0]
		crossover_rate	0.3	[0.0,1.0]

3.5 接受条件 Acceptance Condition

```
<ACCEPTANCE>
    <THRESHOLD>0.98</THRESHOLD>
    <FILE>accepted.nt</FILE>
    <RELATION>owl:sameAs</RELATION>
</ACCEPTANCE>
```

- THRESHOLD: 阈值。
- FILE: 输出文件路径，只支持nt格式。
- RELATION: 实体关系名称。

3.6 复审条件 Review Condition

与Acceptance Condition类似，一般阈值比前者小，某些不满足Acceptance Condition的实体对，可根据Review Condition输出到另一个文件进行复审。

```
<REVIEW>
    <THRESHOLD>0.95</THRESHOLD>
    <FILE>review.nt</FILE>
    <RELATION>owl:sameAs</RELATION>
</REVIEW>
```

3.7 配置文件样例 可按照下面的完整例子进行配置

```
<?xml version="1.0" encoding="UTF-8" standalone="no"?>
<!DOCTYPE LIMES SYSTEM "limes.dtd">
<LIMES>
    <PREFIX>
        <NAMESPACE>http://www.w3.org/2002/07/owl#</NAMESPACE>
        <LABEL>owl</LABEL>
    </PREFIX>
    <PREFIX>
        <NAMESPACE>http://geovocab.org/geometry#</NAMESPACE>
        <LABEL>geom</LABEL>
    </PREFIX>
    <PREFIX>
        <NAMESPACE>http://www.opengis.net/ont/geosparql#</NAMESPACE>
        <LABEL>geos</LABEL>
    </PREFIX>
    <PREFIX>
        <NAMESPACE>http://linkedgeodata.org/ontology/</NAMESPACE>
        <LABEL>lgdo</LABEL>
    </PREFIX>
    <SOURCE>
        <ID>linkedgeodata</ID>
        <ENDPOINT>http://linkedgeodata.org/sparql</ENDPOINT>
        <VAR>?x</VAR>
        <PAGESIZE>2000</PAGESIZE>
        <RESTRICTION>?x a lgdo:RelayBox</RESTRICTION>
        <PROPERTY>geom:geometry/geos:asWKT RENAME polygon</PROPERTY>
    </SOURCE>
    <TARGET>
        <ID>linkedgeodata</ID>
        <ENDPOINT>http://linkedgeodata.org/sparql</ENDPOINT>
        <VAR>?y</VAR>
        <PAGESIZE>2000</PAGESIZE>
        <RESTRICTION>?y a lgdo:RelayBox</RESTRICTION>
        <PROPERTY>geom:geometry/geos:asWKT RENAME polygon</PROPERTY>
    </TARGET>
    <METRIC>geo_hausdorff(x.polygon, y.polygon)</METRIC>
    <ACCEPTANCE>
        <THRESHOLD>0.9</THRESHOLD>
        <FILE>lgd_relaybox_verynear.nt</FILE>
        <RELATION>owl:sameAs</RELATION>
    </ACCEPTANCE>
    <REVIEW>
        <THRESHOLD>0.5</THRESHOLD>
        <FILE>lgd_relaybox_near.nt</FILE>
        <RELATION>owl:sameAs</RELATION>
    </REVIEW>
    <EXECUTION>
        <REWRITER>default</REWRITER>
        <PLANNER>default</PLANNER>
        <ENGINE>default</ENGINE>
    </EXECUTION>

    <OUTPUT>TAB</OUTPUT>
</LIMES>
```

其中EXECUTION和OUTPUT按默认配置。

1. **Metric使用实例**

4.1 余弦(Cosine)相似度比较

4.1.1 数据

Source和Target数据集格式都为nt,均为百科数据的Label信息。 Source数据例子：
`<http://cndbpedia/resource/cdb81e8a-6d56-407a-962a-6d48da6367ad> <http://cndbpedia/ontology/实体名称> "`
Target数据例子：
`<http://zhishi.me/zhwiki/resource/历史> <http://www.w3.org/2000/01/rdf-schema#label> "历史"@zh .`

Source和Target文件均上传至OpenKG.CN:链接

	文件名	实体数	三元组数量
Source	cndbpediaDump_26.nt	358986	927503
Target	zhwiki_labels_zh.nt	575770	575770

4.1.2 配置文件

```
<?xml version="1.0" encoding="UTF-8" standalone="no"?>
<!DOCTYPE LIMES SYSTEM "limes.dtd">
<LIMES>
    <PREFIX>
        <NAMESPACE>http://www.w3.org/2002/07/owl#</NAMESPACE>
        <LABEL>owl</LABEL>
    </PREFIX>
    <PREFIX>
        <NAMESPACE>http://zhishi.me/ontology/</NAMESPACE>
        <LABEL>zhishi</LABEL>
    </PREFIX>
    <PREFIX>
        <NAMESPACE>http://cndbpedia/resource/</NAMESPACE>
        <LABEL>cndb</LABEL>
    </PREFIX>

    <PREFIX>
        <NAMESPACE>http://cndbpedia/ontology/</NAMESPACE>
        <LABEL>cndbo</LABEL>
    </PREFIX>

    <PREFIX>
        <NAMESPACE>http://www.w3.org/2000/01/rdf-schema#</NAMESPACE>
        <LABEL>rdfs</LABEL>
    </PREFIX>


    <SOURCE>
        <ID>cndbpedia</ID>
        <ENDPOINT>cndbpediaDump_26.nt</ENDPOINT>
        <VAR>?x</VAR>
        <PAGESIZE>-1</PAGESIZE>
        <RESTRICTION>?x cndbo:实体名称 ?z1</RESTRICTION>
        <PROPERTY>cndbo:实体名称 RENAME label</PROPERTY>
        <TYPE>NT</TYPE>
    </SOURCE>
    <TARGET>
        <ID>zhwiki</ID>
        <ENDPOINT>zhwiki_labels_zh.nt</ENDPOINT>
        <VAR>?y</VAR>
        <PAGESIZE>-1</PAGESIZE>
        <RESTRICTION>?y rdfs:label ?z2</RESTRICTION>
        <PROPERTY>rdfs:label AS nolang</PROPERTY>
        <TYPE>NT</TYPE>
    </TARGET>
    <METRIC>Cosine(x.label, y.rdfs:label) | 0.8</METRIC>
    <ACCEPTANCE>
        <THRESHOLD>0.9</THRESHOLD>
        <FILE>accept_result.nt</FILE>
        <RELATION>owl:sameAs</RELATION>
    </ACCEPTANCE>
    <REVIEW>
        <THRESHOLD>0.5</THRESHOLD>
        <FILE>review_result.nt</FILE>
        <RELATION>owl:sameAs</RELATION>
    </REVIEW>

    <EXECUTION>
        <REWRITER>default</REWRITER>
        <PLANNER>default</PLANNER>
        <ENGINE>default</ENGINE>
    </EXECUTION>

    <OUTPUT>TAB</OUTPUT>
</LIMES>
```

4.1.3 输出文件样例

```
    <http://cndbpedia/resource/d0fe5706-7a29-4c10-87b3-0882adf33992>    <http://zhishi.me/zhwiki/resou
rce/有机食品> 1.0
    <http://cndbpedia/resource/5efcc63f-4641-477d-bb3e-71d98c382c2a>    <http://zhishi.me/zhwiki/resou
rce/吕龙光>    1.0
    <http://cndbpedia/resource/363a97ee-294b-4de5-b0dc-eb97b6c096c8>    <http://zhishi.me/zhwiki/resou
rce/旺旺>   1.0
```

每一列的数字为度量表达式的计算结果，即相似度。

	Acceptance	**Review**
融合结果数	9517	420041

执行时间：63s

选取Review的10条结果如下：

```
    <http://cndbpedia/resource/"Microsoft. NET Framework 4.0">  <http://zhishi.me/zhwiki/resource/Micr
osoft Windows NT 4.0> 0.6
    <http://cndbpedia/resource/"Microsoft. NET Framework 4.0">  <http://zhishi.me/zhwiki/resource/.NET
 Framework>   0.6324555320336759
    <http://cndbpedia/resource/"Microsoft. NET Framework 4.0">  <http://zhishi.me/zhwiki/resource/Fire
fox 4.0>  0.5163977794943222
    <http://cndbpedia/resource/"Microsoft. NET Framework 4.0">  <http://zhishi.me/zhwiki/resource/.NET
 Compact Framework>
    <http://cndbpedia/resource/"萨里县"> <http://zhishi.me/zhwiki/resource/萨里县 (维吉尼亚州)>  0.7071067
811865475
    <http://cndbpedia/resource/"萨里县"> <http://zhishi.me/zhwiki/resource/萨里县 (北卡罗莱纳州)>   0.7071
067811865475
    <http://cndbpedia/resource/"Premiere Pro">  <http://zhishi.me/zhwiki/resource/Pro Tools>    0.5
    <http://cndbpedia/resource/"Premiere Pro">  <http://zhishi.me/zhwiki/resource/MacBook Pro>  0.5
    <http://cndbpedia/resource/"Premiere Pro">  <http://zhishi.me/zhwiki/resource/Mac Pro>  0.5
    <http://cndbpedia/resource/"Premiere Pro">  <http://zhishi.me/zhwiki/resource/IDA Pro>  0.5
```

可以看到cosine的review匹配效果是比较粗糙的。

4.2 完全匹配(ExactMatch)相似度比较

4.2.1 数据集和配置文件

数据集与上一节相同，配置文件更改METRIC为：

```
    <METRIC>
       ExactMatch(x.label, y.rdfs:label) | 1.0
    </METRIC>
```

4.2.2 结果

	Acceptance	**Review**
融合结果数	9507	0

执行时间：60s

4.3 Cosine 与 ExactMatch比较

ExactMatch是非常严格的相似度比较算法，Review数量为0，准确度Precision很高，但是 召回率Recall就会很低。Cosine则比较均衡，准确度和ExactMatch相似，Review数目很大。

选取Cosine Accept比ExactMatch多的10条结果展示：

```
<http://cndbpedia/resource/"Bonjour!">  <http://zhishi.me/zhwiki/resource/Bonjour>  1.0
<http://cndbpedia/resource/".网络">   <http://zhishi.me/zhwiki/resource/网络>   1.0
<http://cndbpedia/resource/"SHUFFLE MEMORIES">  <http://zhishi.me/zhwiki/resource/SHUFFLE! MEMORIES>
  1.0
<http://cndbpedia/resource/"在路上...">  <http://zhishi.me/zhwiki/resource/在路上>    1.0
<http://cndbpedia/resource/"mobi">  <http://zhishi.me/zhwiki/resource/.mobi>    1.0
<http://cndbpedia/resource/"NO.">   <http://zhishi.me/zhwiki/resource/NO>   1.0
<http://cndbpedia/resource/"va">    <http://zhishi.me/zhwiki/resource/.va>  1.0
<http://cndbpedia/resource/"家庭教师REBORN!">   <http://zhishi.me/zhwiki/resource/家庭教师REBORN>   1.0
<http://cndbpedia/resource/"Godspeed You Black Emperor">    <http://zhishi.me/zhwiki/resource/Godspeed
 You! Black Emperor>  1.0
<http://cndbpedia/resource/".网址">   <http://zhishi.me/zhwiki/resource/网址>   1.0
```

1. **MLALgorithm使用实例**

5.1 领域数据集之间的匹配

5.1.1 数据准备

数据采用分别从PKUPie和Belief-Engine提取的电影领域的数据集。

数据集	实体数	三元组数
pku-movie	15529	33467
belief-movie	4695	35632

其中，belief-movie的4695个实体在pku-movie都有对应的等价实体。

数据下载链接

5.1.2 配置文件

```
<?xml version="1.0" encoding="UTF-8" standalone="no"?>
<!DOCTYPE LIMES SYSTEM "limes.dtd">
<LIMES>
    <PREFIX>
        <NAMESPACE>http://www.w3.org/2002/07/owl#</NAMESPACE>
        <LABEL>owl</LABEL>
    </PREFIX>
    <PREFIX>
        <NAMESPACE>http://www.w3.org/2000/01/rdf-schema#</NAMESPACE>
        <LABEL>rdfs</LABEL>
    </PREFIX>
    <PREFIX>
        <NAMESPACE>http://pkubase/ontology/</NAMESPACE>
        <LABEL>pkuo</LABEL>
    </PREFIX>
    <PREFIX>
        <NAMESPACE>http://www.belief-engine.org/baike_hudong/resource/</NAMESPACE>
        <LABEL>beliefr</LABEL>
    </PREFIX>
    <SOURCE>
        <ID>belief</ID>
        <ENDPOINT>interest_triple_actor_final_belief_label.nt</ENDPOINT>
        <VAR>?x</VAR>
        <PAGESIZE>-1</PAGESIZE>
        <RESTRICTION>?x beliefr:label ?z1</RESTRICTION>
        <PROPERTY>beliefr:label RENAME label</PROPERTY>
        <TYPE>NT</TYPE>
    </SOURCE>
    <TARGET>
        <ID>pku</ID>
        <ENDPOINT>interest_triple_actor_final_pku_label.nt</ENDPOINT>
        <VAR>?y</VAR>
        <PAGESIZE>-1</PAGESIZE>
        <RESTRICTION>?y pkuo:label ?z2</RESTRICTION>
        <PROPERTY>pkuo:label RENAME label</PROPERTY>
        <TYPE>NT</TYPE>
    </TARGET>
    <MLALGORITHM>
        <NAME>wombat simple</NAME>
        <TYPE>supervised batch</TYPE>
        <TRAINING>ml_train_data.nt</TRAINING>
        <PARAMETER>
            <NAME>max execution time in minutes</NAME>
            <VALUE>60</VALUE>
        </PARAMETER>
    </MLALGORITHM>
    <ACCEPTANCE>
        <THRESHOLD>0.9</THRESHOLD>
        <FILE>accept_result.nt</FILE>
        <RELATION>owl:sameAs</RELATION>
    </ACCEPTANCE>
    <REVIEW>
        <THRESHOLD>0.5</THRESHOLD>
        <FILE>review_result.nt</FILE>
        <RELATION>owl:sameAs</RELATION>
    </REVIEW>
    <EXECUTION>
        <REWRITER>default</REWRITER>
        <PLANNER>default</PLANNER>
        <ENGINE>default</ENGINE>
    </EXECUTION>
    <OUTPUT>TAB</OUTPUT>
</LIMES>
```

其中MLALgorithm算法采用wombat simple，训练样本文件格式如下：

```
     <http://www.belief-engine.org/baike_hudong/resource/%E5%BC%A0%E6%81%A9%E5%9F%8E>   <http://www.w3
.org/2002/07/owl#sameAs>  <http://pkubase/entity/957342>
```

保存为.nt文件，并且relation必须是 `<http://www.w3.org/2002/07/owl#sameAs>`

5.1.3 结果

融合结果如下：

训练集三元组	Acceptance	Review
500	4695	0

执行时间：10s

可以得出，采用机器学习算法匹配的结果准确率很高，但是Review为0，wombat simple 不太适合模糊匹配。

5.2 领域与百科数据集之间的匹配

5.2.1 数据准备

数据集分别采用pku-movie和zhishi.me的中文百科数据集(只包含label信息)

数据集	实体数	三元组数
pku-movie	15529	33467
zhishime_wiki_zh	559402	559402

数据下载链接

5.2.2 配置文件

与上一小结的文件类似，不再赘述。

5.2.3 结果

融合结果与ExactMactch(字符串完全匹配)方法做了对比：

训练集三元组	Acceptance	Review	ExactMatch
500	3834	0	3829

执行时间：94s

可以看到Machine Learning的结果跟ExactMatch非常类似，说明wombat simple的准确率是非常高的。

提取出Machine Learning比ExactMatch多匹配到的实体对名称：

zhishi.me	pku-movie
.网络	网络
Hey! Say! JUMP	Hey!Say!JUMP
S.H.E.	S.H.E
巴不得爸爸...	巴不得爸爸
Sound.Horizon	Sound Horizon

可以看到，单纯使用ExactMatch是无法提取出这些实体对的，但是Machine Learning找到了，并且经过人工判断，也是正确的。