利用关联规则在结构化三元组中查找相同语义的谓词

Synonym Analysis for Predicate Expansion

eswc2013

Ziawasch Abedjan @hpi.de 哈索·普拉特纳研究院

Felix Naumann@hpi.de

hanzhe@icst-wip

2016 年 1 月 22 日

Outline

- summary
  - 论文动机
    - 找到 LOD（linked-open-data）里面重复的谓词
  - 方法概述
    - 将谓词加到关联规则中，挖掘相同的谓词对
  - dbpedia 实验效果

关联规则

- 项集 (item set)： $I = \{i_1, i_2, \ldots\ i_m\}$
- 事务集 (Transaction set)： $T = \{t | t \subseteq I\}$
- 一个关联规则 $X \rightarrow Y$，其中 $X, Y \subseteq I, X \cap Y = \emptyset$
  - 该关联规则的支持度 (support): $\frac{|\{t|t \in T, X \cup Y \subseteq t\}|}{|T|}$
  - 该关联规则的置信度 (confidence): $\frac{|\{t|t \in T, X \cup Y \subseteq t\}|}{|\{t|t \in T, X \subseteq t\}|}$

关联规则

Table 2: Facts in SPO structure from DBpedia

Subject	Predicate	Object
Obama	birthPlace	Hawaii
Obama	party	Democrats
Obama	orderInOffice	President
Merkel	birthPlace	Hamburg
Merkel	orderInOffice	Chancellor
Merkel	party	CDU
Brahms	born	Hamburg
Brahms	type	Musician

Table 3: Six configurations of context and target

Conf.	Context	Target	Use case
1	Subject	Predicate	Schema discovery
2	Subject	Object	Basket analysis
3	Predicate	Subject	Clustering
4	Predicate	Object	Range discovery
5	Object	Subject	Topical clustering
6	Object	Predicate	Schema matching

TID	transaction
Obama	{*birthPlace, party, orderInOffice*}
Merkel	{*birthPlace, party, orderInOffice*}
Lennon	{*birthPlace, instrument*}

- $birthPlace \rightarrow orderInOffice$ 置信度 66.7%，支持度 66.7%，$orderInOffice \rightarrow birhtPlace$ 置信度 100%，支持度 66.7%

方法说明

—

方法说明

如果谓词 p_1, p_2 意义相同，则：

1. 他们不会出现在同一个主语的谓词集合里
2. 他们所在的三元组含有很多相同的客体
3. 他们所在的三元组的客体类别的分布很相似

方法说明

如果谓词 p_1, p_2 意义相同

1. 他们不会出现在同一个主语的谓词集合里（RCC）

Table 2: Facts in SPO structure from DBpedia

Subject	Predicate	Object
Obama	birthPlace	Hawaii
Obama	party	Democrats
Obama	orderInOffice	President
Merkel	birthPlace	Hamburg
Merkel	orderInOffice	Chancellor
Merkel	party	CDU
Brahms	born	Hamburg
Brahms	type	Musician

Table 3: Six configurations of context and target

Conf.	Context	Target	Use case
1	Subject	Predicate	Schema discovery
2	Subject	Object	Basket analysis
3	Predicate	Subject	Clustering
4	Predicate	Object	Range discovery
5	Object	Subject	Topical clustering
6	Object	Predicate	Schema matching

TID	transaction
Obama	{*birthPlace, party, orderInOffice*}
Merkel	{*birthPlace, party, orderInOffice*}
Lennon	{*birthPlace, instrument*}

- $X \to \neg Y, Y \to \neg X$ 置信度都很高（避免 Y 出现频率很低造成 $X \to \neg Y$）
- 比如 $party \to \neg instrument$,

  $birthPlace \to \neg born$

方法说明

如果谓词 p_1, p_2 意义相同

1. 他们不会出现在同一个主语的谓词集合里（RCC）

Table 2: Facts in SPO structure from DBpedia

Subject	Predicate	Object
Obama	birthPlace	Hawaii
Obama	party	Democrats
Obama	orderInOffice	President
Merkel	birthPlace	Hamburg
Merkel	orderInOffice	Chancellor
Merkel	party	CDU
Brahms	born	Hamburg
Brahms	type	Musician

Table 3: Six configurations of context and target

Conf.	Context	Target	Use case
1	Subject	Predicate	Schema discovery
2	Subject	Object	Basket analysis
3	Predicate	Subject	Clustering
4	Predicate	Object	Range discovery
5	Object	Subject	Topical clustering
6	Object	Predicate	Schema matching

TID	transaction
Obama	{*birthPlace, party, orderInOffice*}
Merkel	{*birthPlace, party, orderInOffice*}
Lennon	{*birthPlace, instrument*}

- $X \to \neg Y, Y \to \neg X$ 置信度都很高（避免 Y 出现频率很低造成 $X \to \neg Y$）

$$cCoeff(X,Y) = \frac{N \cdot supp(X,Y) - supp(X) \cdot supp(Y)}{\sqrt{supp(Y) \cdot (N - supp(Y)) \cdot supp(X) \cdot (N - supp(X))}}$$

方法说明

如果谓词 p_1, p_2 意义相同

2. 他们所在的三元组含有很多相同的客体（RCF-range content filtering）

Table 2: Facts in SPO structure from DBpedia

Subject	Predicate	Object
Obama	birthPlace	Hawaii
Obama	party	Democrats
Obama	orderInOffice	President
Merkel	birthPlace	Hamburg
Merkel	orderInOffice	Chancellor
Merkel	party	CDU
Brahms	born	Hamburg
Brahms	type	Musician

Table 3: Six configurations of context and target

Conf.	Context	Target	Use case
1	Subject	Predicate	Schema discovery
2	Subject	Object	Basket analysis
3	Predicate	Subject	Clustering
4	Predicate	Object	Range discovery
5	Object	Subject	Topical clustering
6	Object	Predicate	Schema matching

TID	transaction
Musician	$\{type\}$
Hamburg	$\{born, birthPlace\}$
Hawaii	$\{birthPlace\}$
President	$\{orderInOffice\}$

- *born* 和 *birthPlace*
- 比如 $party \rightarrow \neg instrument$,

  $birthPlace \rightarrow \neg born$

方法说明

如果谓词 p_1, p_2 意义相同

3. 他们所在的三元组含有很多相同的客体（RSF-range Structure/type filtering）

类似 range content filtering，将 content 用对应类别替代，计算任意一个谓词对 p_1, p_2，只需要计算类别向量的相似度

方法说明

如果谓词 p_1, p_2 意义相同，则前面 3 种想法混合：

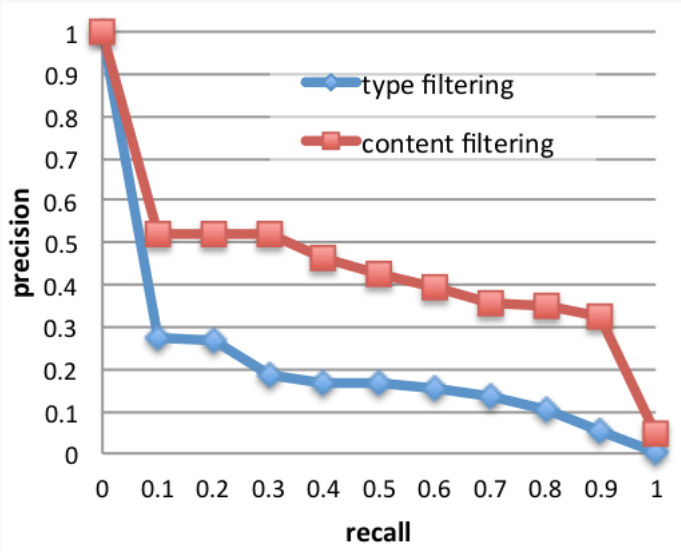
1. 根据（RCF-range content filtering）得到所有的谓词对候选（每对谓词至少重复一个客体）
2. 根据（RSF-range Structure/type filtering）进一步筛选谓词对 (每对谓词至少有一个相同的客体类别)
   - 第一阶段可以重复的客体可能是数值或时间（没有类别），不算相同的可以类别
3. 使用不同的评价标注（RCC/minConf/maxConf/...）计算每个谓词对的相似度

实验

—

实验一

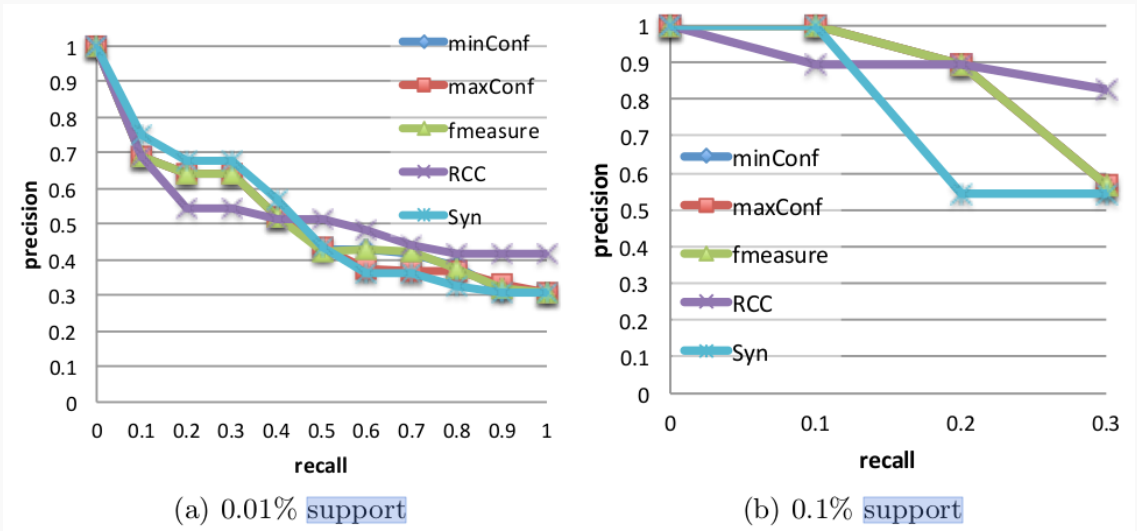
- 比较使用客体类别向量/客体值来判断为此对相似性



实验二

- syn 是别人的方法：相同的谓词不共现、含有相似客体
- minConf, maxConf, fmeasure 与 RCC 类似
- 希望利用“$X \rightarrow \neg Y$ 是一个频繁模式”来表达两个相同的谓词
- minConf=$min\{conf(X \rightarrow \neg Y), conf(Y \rightarrow \neg X)\}$
- maxCont=$max\{...\}$
- fmeasur 是 $minCont$ 和 $maxCont$ 的调和平均数
- RCC=

抽取在 dbpedia ‘Work‘(作品) 类别下的 9456 个谓词对，其中 82 对意思相同



(a) 0.01% support (b) 0.1% support

$$cCoeff(X,Y) = \frac{N \cdot supp(X,Y) - supp(X) \cdot supp(Y)}{\sqrt{supp(Y) \cdot (N - supp(Y)) \cdot supp(X) \cdot (N - supp(X))}}$$

实验三

通过 RCF 对谓词对进行过滤的效果

Table 6: Precision at 0.01% RCF minimum support

Dataset	minConf	maxConf	f-Measure	RCC	Syn	RCF	# RCF results
Magnatune	**100%**	87.5%	**100%**	**100%**	87.5%	87.5%	8
Govwild	0%	**20%**	0%	14%	0%	20%	25
DBpedia 3.7	**32%**	**32%**	**32%**	15%	22%	32%	1115
DBpedia Person	32%	32%	32%	**35%**	26%	32%	308
DBpedia Work	49%	52%	50%	**61%**	60%	22%	256
DBpedia Organisation	**33%**	32%	32%	31%	32%	32%	412

Table 7: Precision values at 0.1% range content filtering minimum support

Dataset	minConf	maxConf	fMeasure	RCC	Syn	RCF	# RCF results
Magnatune	**100%**	**100%**	**100%**	**100%**	**100%**	100%	4
Govwild	0%	**56%**	0%	50%	0%	50%	10
DBpedia 3.7	40%	43%	38%	**46%**	45%	36%	64
DBpedia Person	56%	49%	50%	**60%**	-	40%	35
DBpedia Work	73%	57%	74%	78%	**89%**	52%	46
DBpedia Organisation	88%	86%	90%	89%	**95%**	85%	45

- 上下两个图对比，可以发现 RCf 过滤有用
- 上图的 Dbpedia work 数据集 49% 的准确率比实验二的 30% 左右的准确率高出不少

谢谢大家