Vysoká škola ekonomická v Praze
Fakulta informatiky a statistiky

Minireferat na
4IZ440 Propojená data na webu
„CubeQA - Question Answering on RDF Data Cubes“

Bc. Danil Votinov
ZS 2016/2017

Úvod

Statistické údaje v podobě RDF Data Cubes se stává cennější v rozhodnutích, jak to ovlivňuje rozhodování v oblastechjako jako zdravotní péče, politika a financí. SQA technologie poskytují intuitivní přístup prostřednictvím volného tvaru přirozeného jazyka dotazů ale obecné SQA systémy nemohou zpracovávat RDF Data Cubes. Křižovatka mezi SQA a RDF Data Cubes - nove podpole SQA - RDCQA.

Na rozdíl od společné reprezentace dat, jako tabulky nebo relační databázi Data Cubovy formalismus představuje multidimenzionální, číselné údaje. Data Cude je vícerozměrné řada buněk. Každá buňka je jednoznačně identifikována jeho dimenze a obsahuje jednu nebo více číselnu miru. Data cubes jsou často „sparse“. Data Cubes umožňují následující operace:

1. Nasekat Data Cube (Dicing) - vytváří SubCube omezováním Cubu do podmnožiny svých hodnot.
2. Krájení Data Cubu (Slicing) - snižuje jeho rozměrnost omezováním dimenze k jedné konkrétní hodnote.
3. Vyhrnout (Rolling Up) - shrnující mir po dimenze (součet, počet, aritmetický průměr).

Definice 1. Odpověď na otázku (Question answering, QA) - Uživatel (1) položí otázku v přirozeném jazyce (natural language), (2) za použití jejich vlastní terminologii, kterou (3) obdrží stručnou odpověď. V SQA otázka v přirozeném jazyce transformuje do formální dotazu na znalostní bázi RDF (běžně používame SPARQL).

RDF Data Cude slovnikovy modely zahrnují schéma a pozorování Dada Cubu do RDF. Kazdy Data Cude (instancí *qb:DataSet*) ma připojenou schématu - definice datovou struktury (data structure definition) ktera určuje vlastnosti komponenty (component properties). Vlastnost komponenty (component property) je dimenze a take atribut miry. Musí existovat alespoň jedna mira, ktera představují měřené veličiny, zatímco dimenze a volitelné atributy poskytnuji kontext.

Kombinování Matches v Constraints (omezeni).

Definice 2. Match m představuje formálně dvojice (p, y) kde p je částečna vlastnost vyhodnocováne funkce, y je částečna hodnota vyhodnocováne funkce.

Rekurzivní kombinační proces se používá, protože zvýhodňuje delší frází před menšími, dáva větší pokrytí dané otázce a zvýhodňuje kombinaci frází, které jsou poblíž v derivačním stromu otázce.

Definice 3. Omezení(Constraint) c je reprezentována trojice (G, w, λ), kde

- G je množina SPARQL vzorovych trojic a filtru;
- w je volitelný pořádek podle modifikátoru;
- λ je volitelný modifikátor limitu.

Omezení jsou založeny na kritérií:

1. Hodnota omezení může být aplikována na jakoukoliv vlastnost komponentu , aby jej omezit na přesnou hodnotu, která může být řetězec, číslo nebo URI;
2. Interval omezení omezuje hodnotu na číselný nebo časový interval.

Algoritmus. Fragment k obrácení šablony

Input: Fragment $f = \{(M, R)\}$ volitelná agregační funkce α, ktera identifikována v předzpracování.
Output: šablona $t = (R', a, \alpha')$.
$R' = R$;
$P' \leftarrow \big(P \backslash A \backslash \delta(R)\big)$;
$\boldsymbol{foreach}\ (\rho, \gamma) \in M' \boldsymbol{do}\ \{$
$\quad p_{max} = arg\ max_{p \in (dom(\gamma) \cap P')} \big(\pi_2(\gamma(p))\big)$;
$\quad \boldsymbol{if}\ \Big((p_{max} \neq null) \wedge \big(\pi_2(\gamma(p_{max}))\big) \leq \theta\Big)\ \boldsymbol{then}$
$\quad\quad R' \leftarrow R' \cup \big(\{?\, o\ p_{max}\ \pi_1(\gamma(p_{max}))\, .\}, null, null\big)$;
}
$A \leftarrow \cup_{(\rho,\gamma) \in M}\ dom(\rho)$;
$A' \leftarrow \{p \in A | answerType(p) \in E\}$;
$\boldsymbol{if}\ A' = 0\ \boldsymbol{then}$
$\quad a \leftarrow DEFAULT_MEASURE$;
$\boldsymbol{else}$
$\quad a \leftarrow \arg max_{(\rho,\gamma) \in M, p \in A'} \rho(p)$;
$\boldsymbol{return}\ (R', a, \alpha)$

QALD6T3 - test

QALD6T3 poskytuje několik problémů, které jsou podporovány algoritmu CubeQA. Ty jsou naznačeny agregace, intervaly nebo jinak odkazované miry a číselné hodnoty, které jsou obsaženy v několika vlastnostech komponent. To také zahrnuje otázky, které vyžadují funkce, které nejsou poskytované CubeQA, například SPARQL poddotazů. Výkon CubeQA na benchmarku se měří takto: C - správný soubor prostředků a O - výstup algoritmu. Přesnost (precision) $p = \frac{|C| \cap |O|}{|O|}$, odvolání (recall) $r = \frac{|C| \cap |O|}{|C|}$ a F_1 - score $F_1 = 2\frac{pr}{p+r}$. Průměrné globální skóre F_1 počítá $p = 0$ pro prázdné odpovědi.

Mapování m za tázací slovo na množinu očekávaných typů odpovědí E, spolu s frekvencí každeho tázaceho slova v benchmarku. Když je tázací slovo není známe nebo nebylo nalezeno, vše 6 typů odpovědi jsou možné.

Tázací slovo	Očekávaný typ odpovědí
What	Nepočitatelný, počitatelný, počítat, temporální, umístění, entita
How much	Nepočitatelný
Which	Temporální, umístění, entita
How many	Počitatelný, počítat

When	Temporální
None, other	Nepočitatelný, počitatelný, počítat, temporální, umístění, entita

Závěr

CubeQA bude silným výchozí v novém výzkumném podpole. QALD6T3 je zahájena před předložením publikaci další abstraktní výzkum.

Reference

Höffner, K., Lehmann, J., Usbeck, R. : CubeQA - Question Answering on RDF Data Cubes – The Semantic Web – ISWC 2016 – Kobe, 2016. – pp. 325-340.