Incorporating Information Extraction in Database Queries

Yoav Nahshon, Liat Peterfreund, Stijn Vansummeren, WebDB16 (Best Paper)
Advisor: Prof. Benny Kimelfeld

Technion
Israel Institute of Technology

!"#$%&'($")*+,%'-($")

Information extraction (IE) is the identification, and consequent or concurrent classification and structuring into semantic classes, of specific information found in unstructured data sources, such as natural language text, making the information more suitable for information processing task."

WO'XO'W2/-4$'M/D27V6<2/'YS576;<2/Z'JR12705>V4'6/9'[724K-;54'0/'6'\-57

''9656'0/'5 9656'0/''9656864-'

3"4#$%&)5$601)

=>-'-./0/1'45675-9'G05>'5>-'8-45'
6KK-<L-74'5>65'MN.-'-.-7'5645-90'=>-/'
05'-./'125'8-P-7Q'=>-'76.02R0'G64'
-S;-RR-/5'6/9'5>-'9-44-754'G-7-'
0/;7-908R-QQQ'

BK6//-74'

7$8"))	96:0-(;0)
!"#$%#&'	!"($"%&'
!))$*(+&'	!***$*,,&'

4K6/'

=>-'-./0/1'45675-9'G05>'5>-'8-45'
6KK-<L-74'5>65'MN.-'-.-7'5645-90'
=>-/'05'-./'125'8-P-7Q'=>-'
76.02R0'G64'-S;-RR-/5'6/9'5>-'
9-44-754'G-7-'0/;7-908R-QQQ'

F]['

./"")	<0&&')	=>.)
!#$*"&'	-./0/1'	/23/'
!*#$,"&'	45675'	.-78'
!"($$"#&'	1229'	69:-;<.-'

!6)	?0@,'8%"',)
*,"'	=>-'?-45'?6@-
%#+'	B3/4>0/-'C6DE

!6)	A4B)
%#+'	F-G'H27@'
%#I'	J176'

B_]'

?0@,'8%"',)	A4B)
B3/4>0/-'C6DE'	F-G'H27@'

!6)	?0;40C)
*",'	=>-'-./0/1'45675-9'G05>'5>-'8-45'6KK-<L-74'5>65'MN.-'-.-7'5645-90'=>-/'05'-./'125'8-P-7Q'=>-'76.02R0'G64'-S;-RR-/5'6/9'5>-'9-44-754'G-7-'0/;7-908R-QQQ'

!6)	A4B)
*",'	^60D6'
""%'	=-R'.0.'

'/0D27V'

A4B)	?0;40C)
^60D6'	=>-'-./0/1'45675-9'G05>'5>-'8-45'6KK-<L-74'5>65'MN.-'-.-7'5645-90'=>-/'05'-./'125'8-P-7Q'=>-'76.02R0'G64'-S;-RR-/5'6/9'5>-'9-44-754'G-7-'0/;7-908R-QQQ'

./'""'0%'$2)))

Spannerlog is a declarative language for information extraction that extends Datalog. It defines relations both over strings and spans and therefore enables to reason about the structure of a string as well as its content. The language incorporates different kinds of IE functions such as natural language processing functions and spanners.

=>-';2T--'G64'6GD3R'
835'5>-'K6/;6@-'G64'
9-U/05-RA'5>-'8-45'
K6/;6@-'M'>6.-'-.-7'
-65-/'

Dependency

DEPb(c'+,)*-,.*

>Q0-,)	96:0-(;0)
!#$*,&'	!*I$,,&'
!"*$"a&'	!%"$#%&'

=>-';2T--'G64'6GD3R'
835'5>-'K6/;6@-'G64'
9-U/05-RA'5>-'8-45'
K6/;6@-'M'>6.-'-.-7'
-65-/'

Sentiment Analysis

SAb(8-,9c',67.*'

.0"(&0",)
/-16<.-'
K240<.-'

Program Input:

IE functions can operate on substrings

%0;40C	46
=>-'-./0/1'45675-9'G05>'5>-'8-45'6KK-<L-74'5>65'MN.-'-.-7'5645-90'=>-/'05'-./'125'8-P-7Q'=>-'76.02R0'G64'-S;-RR-/5'6/9'5>-'9-44-754'G-7-'0/;7-908R-QQQ'	
'=>-';2T--'G64'6GD3R'835'5>-'K6/;6@-'G64'9-U/05-RA'5>-'8-45'K6/;6@-'M'>6.-'-.-7'-65-/'	

Program Rules:

EDBs = IE functions

!"#$%&'()+,)*-,.*/*012"13,'()*+,)*-,.**

!1$561$7'(8+,9),67.*/*!"#$%&'()*+,)*-,.(8-,9)',67.***

Intermediate relation:

.42"'1!

%0;40C	$@	:@
=>-';2T--'G64'6GD3R'835'5>-'K6/;6@-'G64'9-U/05-RA'5>-'8-45'K6/;6@-'M'>6.-'-.-7'-65-/'	!#$*,&'	!*I$,,&'
=>-'-./0/1'45675-9'G05>'5>-'8-45'6KK-<L-74'5>65'MN.-'-.-7'5645-90'=>-/'05'-./'125'8-P-7Q'=>-'76.02R0'G64'-S;-RR-/5'6/9'5>-'9-44-754'G-7-'0/;7-908R-QQQ'	!"+$%I&'	!"($"#&'

Program Output:

.0"(&0",!)	
28:-;5	4V5
;2T--'	/-16<.-'
K6/;6@-'	K240<.-'
6KK-<L-74'	K240<.-'
76.02R0'	K240<.-'
9-4-754'	K240<.-'