

Update-Aware Information Extraction

Besat Kassaie

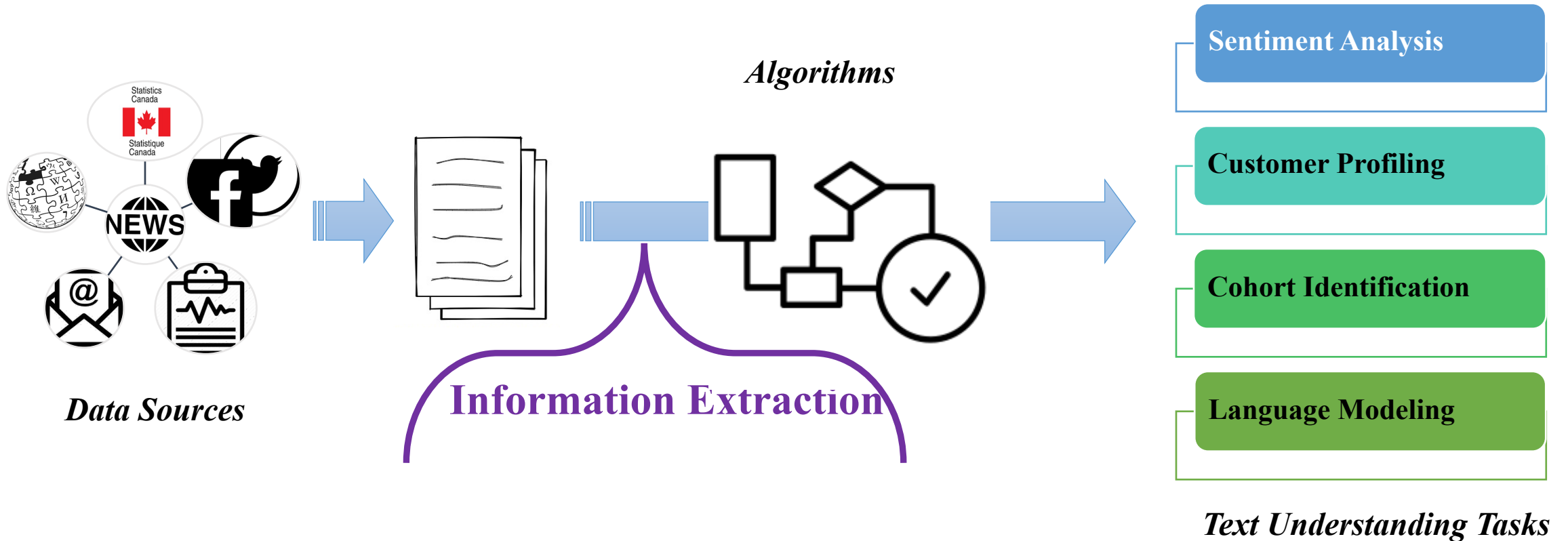


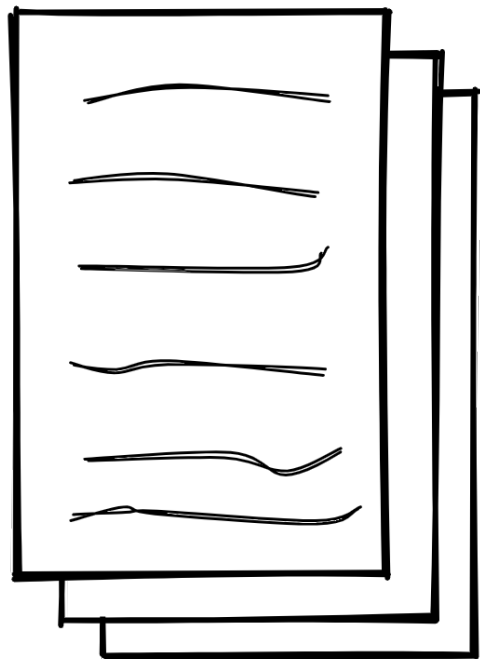
UNIVERSITY OF
WATERLOO



Data
Systems
Group

Text is everywhere ...





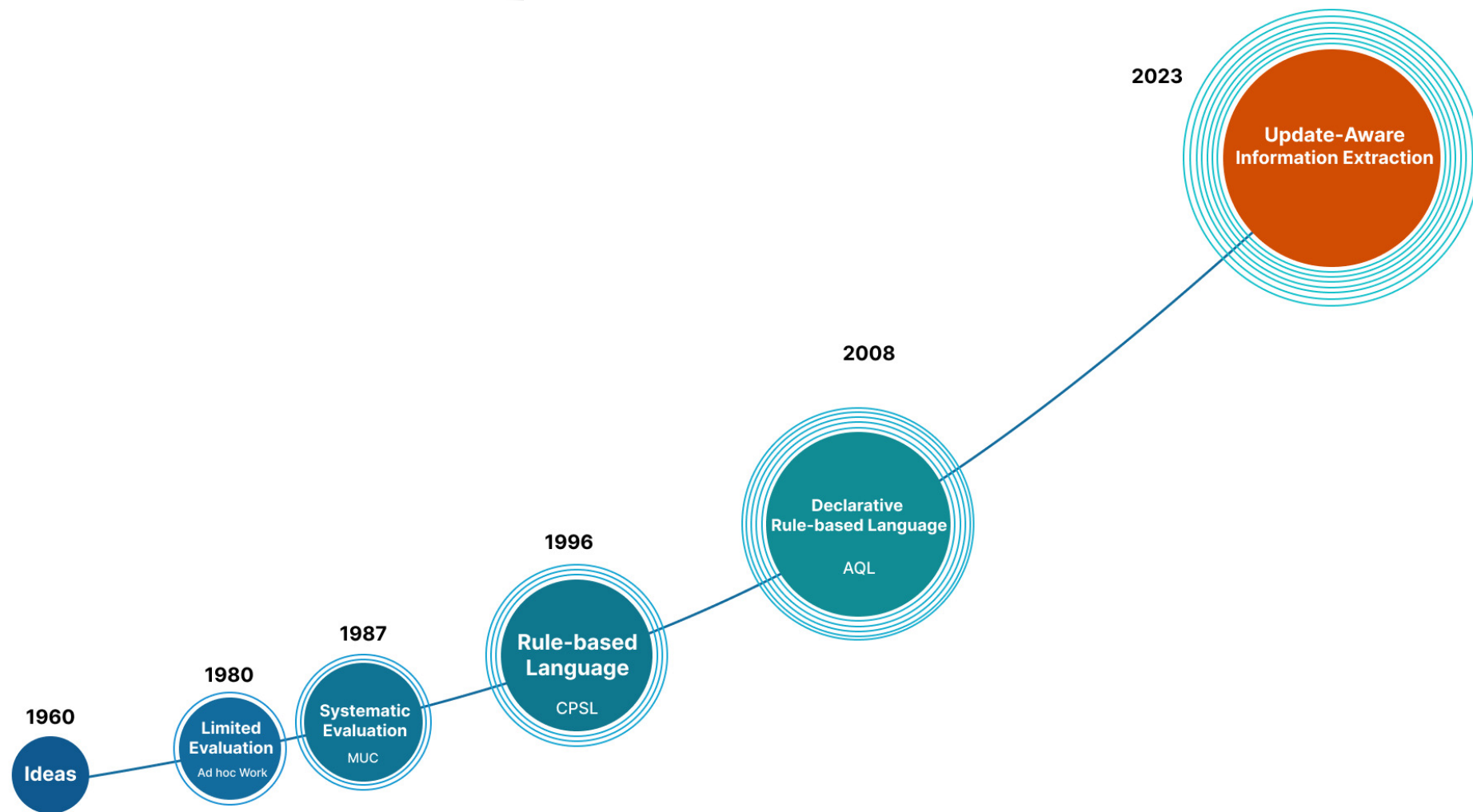
Documents



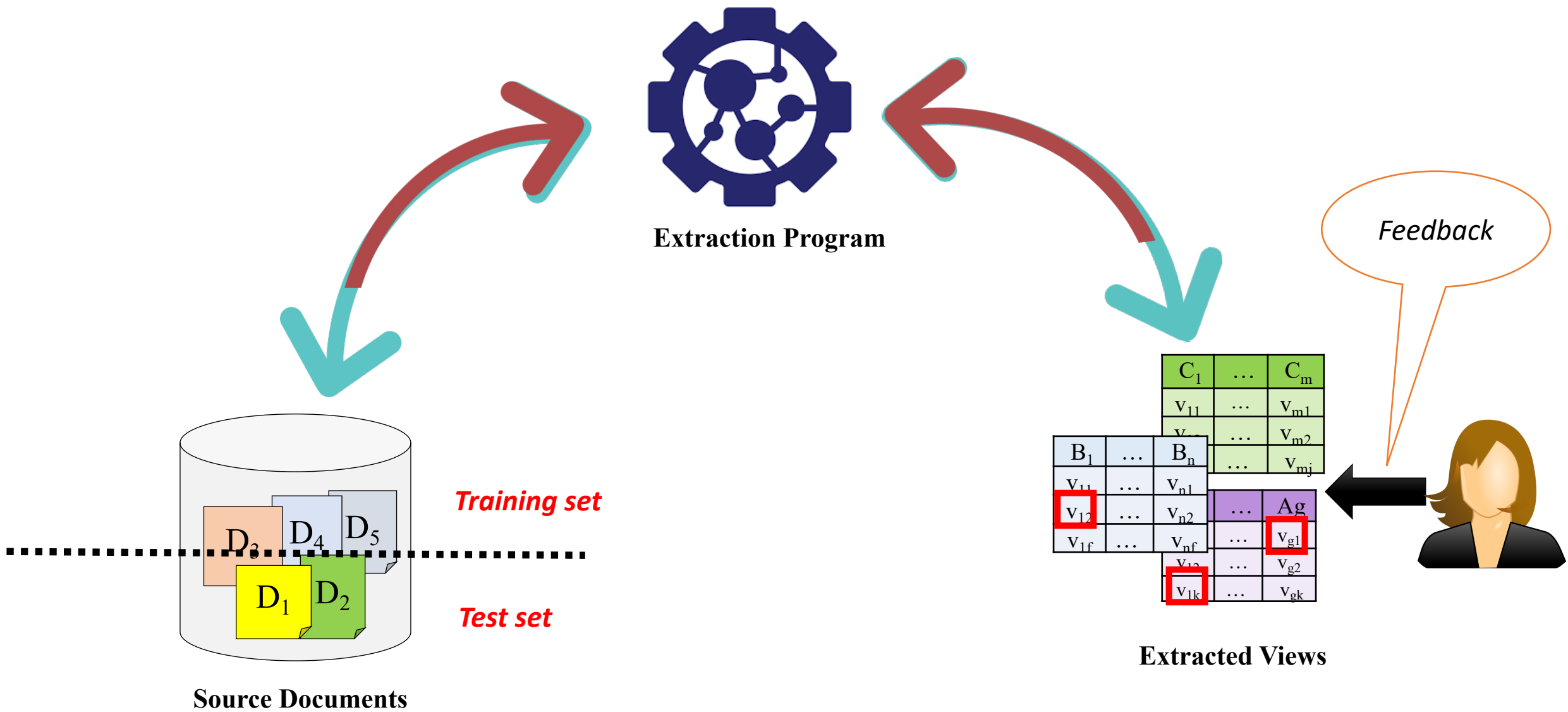
DocID	Name	Address

Extracted Relation

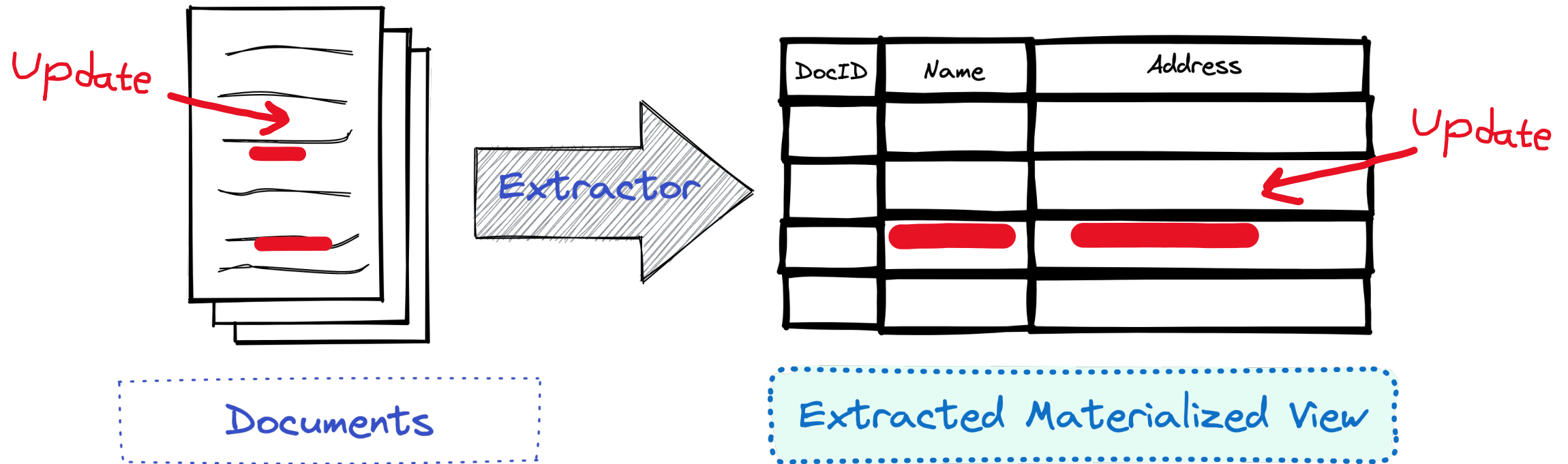
History



Fading Attachment



Main Contribution: Extracted Relations as Materialized Views



Research Problems

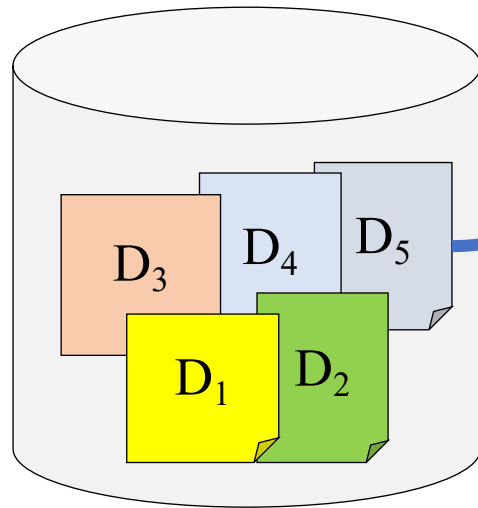
```
graph TD; A[Research Problems] --- B[Updatable Extracted View]; A --- C[Extracted View Maintenance]; C --- D[Today's Talk]
```

Updatable Extracted View

Extracted View Maintenance

Today's Talk

On_03/24, _we_ r e n t e d _ a n d _ w a t c h e d _ ' A _ M a n ' . _
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41
I _ h i g h l y _ r e c o m m e n d _ i t _ a s _ a n _ i n s p i r i n g _ a n d
42 43 44 45 46 47 48 49 50 51 52 53 54 55 56 57 58 59 60 61 62 63 64 65 66 67 68 69 70 71 72 73 74 75 76 77 78 79 80 81 82
_ h u m o r o u s _ f i l m _ t o _ e n j o y .
83 84 85 86 87 88 89 90 91 92 93 94 95 96 97 98 99 100 101 102 103 104 105 106



Source Documents

Document-at-a-Time Extraction

On_03/24, _we_ r e n t e d _ a n d _ w a t c h e d _ ‘ A _ M a n ’ . _

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41

I _ h i g h l y _ r e c o m m e n d _ i t _ a s _ a n _ i n s p i r i n g _ a n d

42 43 44 45 46 47 48 49 50 51 52 53 54 55 56 57 58 59 60 61 62 63 64 65 66 67 68 69 70 71 72 73 74 75 76 77 78 79 80 81 82

_ h u m o r o u s _ f i l m _ t o _ e n j o y .

83 84 85 86 87 88 89 90 91 92 93 94 95 96 97 98 99 100 101 102 103 104 105 106

EXTRACTOR: Identify and extract all occurrences of
words, attributing qualities to some
movies.

Extract Offsets

On **03/24**, we rented and watched 'A Man' .
 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41
 I highly recommend it as an inspiring and
 42 43 44 45 46 47 48 49 50 51 52 53 54 55 56 57 58 59 60 61 62 63 64 65 66 67 68 69 70 71 72 73 74 75 76 77 78 79 80 81 82
humorous film to enjoy .
 83 84 85 86 87 88 89 90 91 92 93 94 95 96 97 98 99 100 101 102 103 104 105 106

EXTRACTOR: Extract all occurrences of words,
 attributing qualities to movies.

A	T
[70, 79>	[93, 97>
[84, 92>	[93, 97>

Span relation

Span

On_03/24, _we_ r e n t e d _ a n d _ w a t c h e d _ ‘ A _ M a n ’ . _
 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41
 I _ h i g h l y _ r e c o m m e n d _ i t _ a s _ a n _ i n s p i r i n g _ a n d
 42 43 44 45 46 47 48 49 50 51 52 53 54 55 56 57 58 59 60 61 62 63 64 65 66 67 68 69 70 71 72 73 74 75 76 77 78 79 80 81 82
 _ h u m o r o u s _ f i l m _ t o _ e n j o y .
 83 84 85 86 87 88 89 90 91 92 93 94 95 96 97 98 99 100 101 102 103 104 105 106

EXTRACTOR: Extract all occurrences of words,
 attributing qualities to movies.

[70, 79>	[93, 97>
[84, 92>	[93, 97>

UPDATE: In each document, re-order
Month and Day and insert
2023/ at the beginning of
 each instance.

F	M	D
[4, 9>	[4, 6>	[7, 9>

On_2023/24/03, _we_ r e n t e d _a n d _w a t c h e d _ ' A _ M
 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41
 a n ' . _ I _ h i g h l y _ r e c o m m e n d _ i t _ a s _ a n _ i n s p i r i n
 42 43 44 45 46 47 48 49 50 51 52 53 54 55 56 57 58 59 60 61 62 63 64 65 66 67 68 69 70 71 72 73 74 75 76 77 78 79 80 81 82
 g _ a n d _ h u m o r o u s _ f i l m _ t o _ e n j o y .
 83 84 85 86 87 88 89 90 91 92 93 94 95 96 97 98 99 100 101 102 103 104 105 106 107 108 109 110 111

EXTRACTION

Extract all occurrences of words, attributing qualities to movies.

A	T
[70, 79>	[93, 97>
[84, 92>	[93, 97>

Inconsistent

On **2023/24/03**, we rented and watched 'A_M

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41

an'. I highly recommend it as an inspiring

42 43 44 45 46 47 48 49 50 51 52 53 54 55 56 57 58 59 60 61 62 63 64 65 66 67 68 69 70 71 72 73 74 75 76 77 78 79 80 81 82

g_and_humorous_film_to_enjoy.

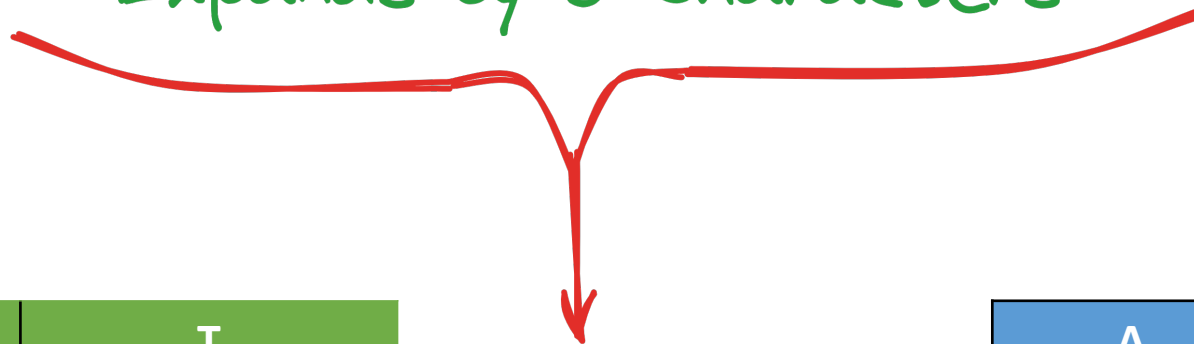
83 84 85 86 87 88 89 90 91 92 93 94 95 96 97 98 99 100 101 102 103 104 105 106 107 108 109 110 111

03/24



2023/24/03

Expands by 5 Characters

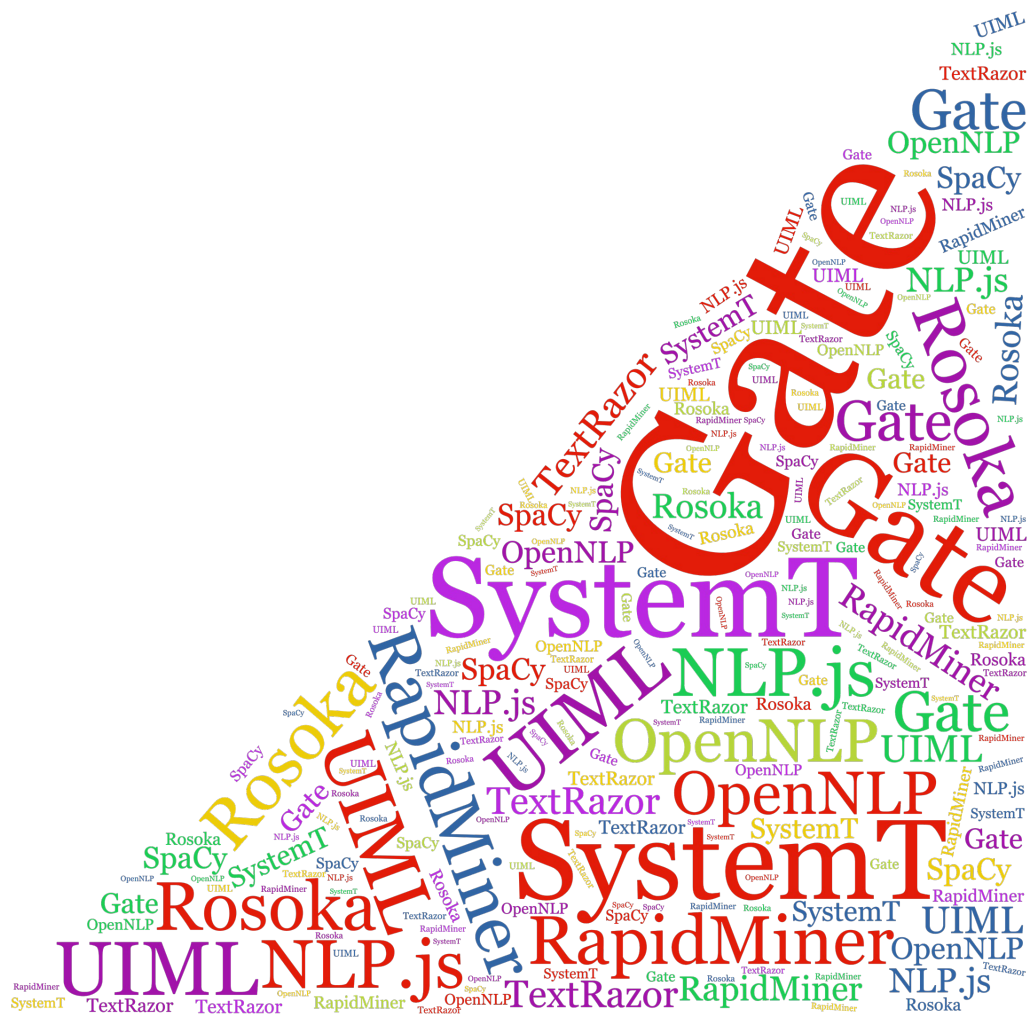


A	T
[70, 79>	[93, 97>
[84, 92>	[93, 97>



Shift

A	T
[75, 84>	[98, 102>
[89, 97>	[98, 102>



Extractors



Strict

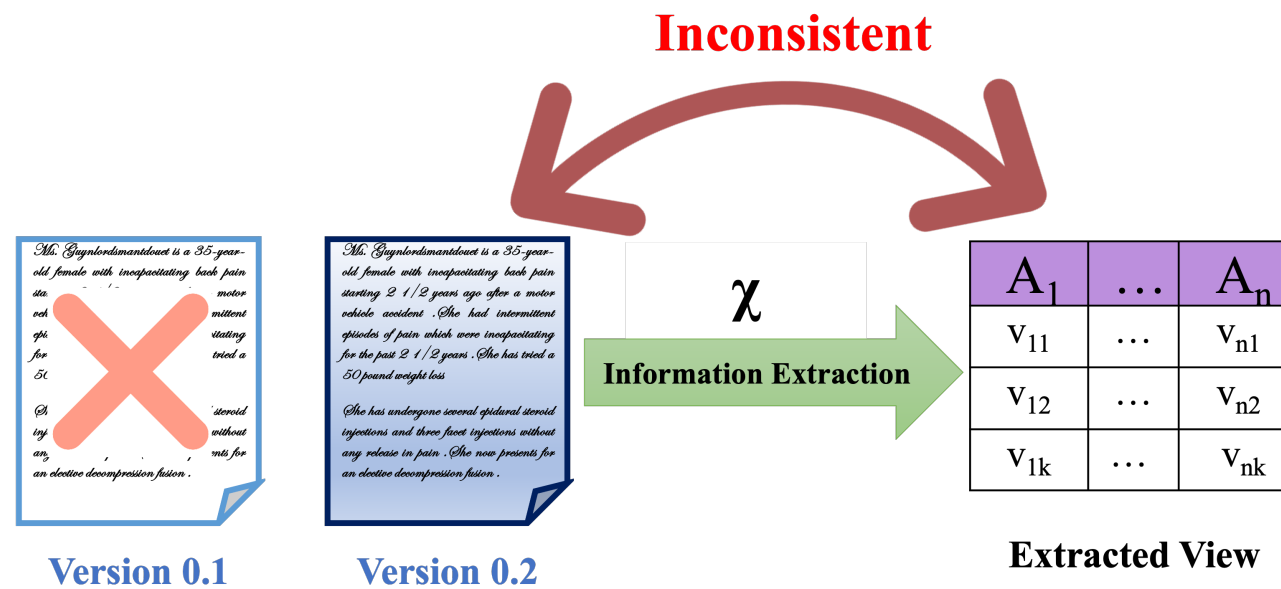
Extracted values appear in the input document.

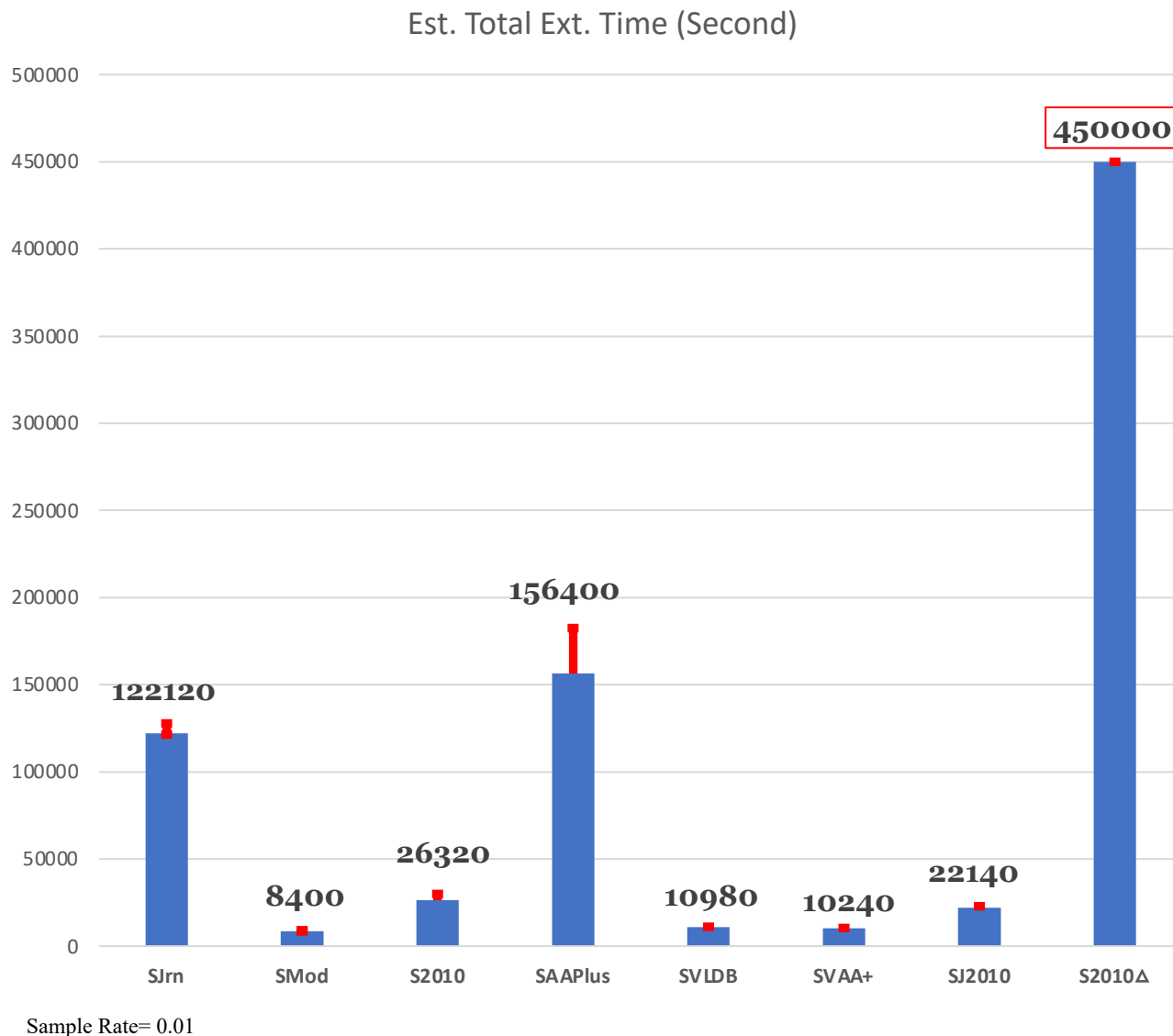
Computable

Positions of extracted values are available.

Rule-based Information Extraction Systems

Extracted View Maintenance



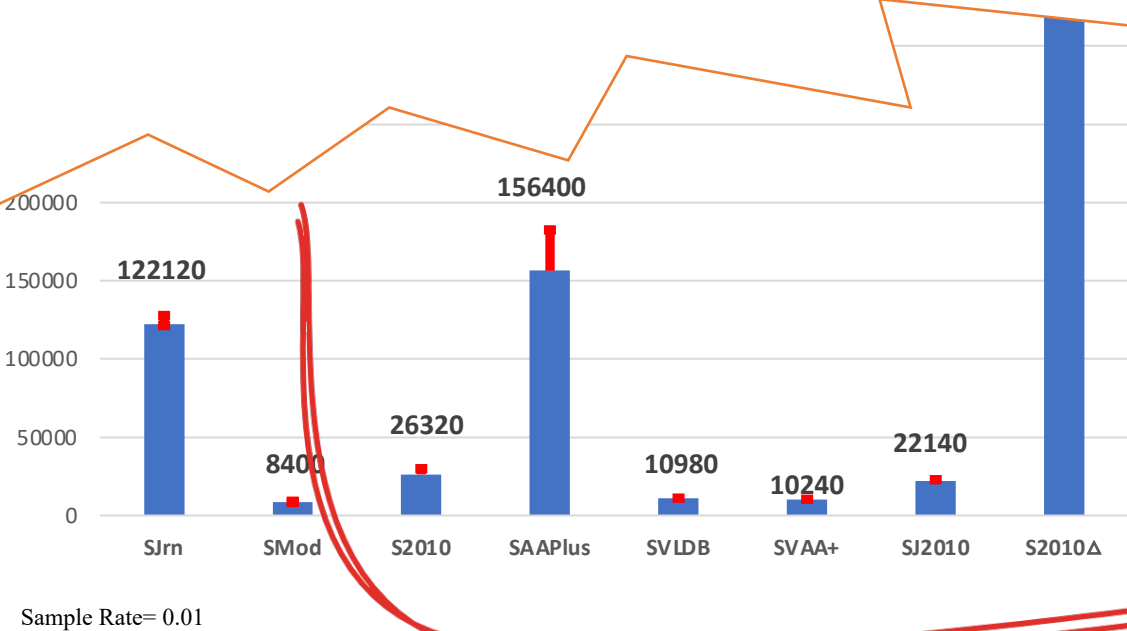


SJrn	Ext. article title, pages, year, volume, and journal from journal articles
SMod	Ext. titles of publications last modified in 2015
SAA+	Ext. first authors of pubs with 2+ authors
S2010	Ext. titles of publications from 2010
SVLDB	Ext. first authors of publications from VLDB
SVAA+	Ext. multi-author VLDB articles
SJ2020	Ext. journal titles from 2010
S2010Δ	Ext. titles from 2010 or updated in 2015

Extraction is slow!

Corpus: [DBLP](#) Number of documents: [3,091,270](#) Corpus size: [7.9 GB](#) Threaded: [Single threaded](#) Machine: [Ubuntu 20.04.1 LTS \(Focal Fossa\)](#) and [AMD EPYC 7502P 32-Core Processor](#).

FREQUENT UPDATES



SJrn	Ext. article title, pages, year, volume, and journal from journal articles
SMod	Ext. titles of publications last modified in 2015
SAA+	Ext. first authors of pubs with 2+ authors
S2010	Ext. titles of publications from 2010
SVLDB	Ext. first authors of publications from VLDB
SVAA+	Ext. multi-author VLDB articles
SJ2020	Ext. journal titles from 2010
S2010Δ	Ext. titles from 2010 or updated in 2015

Extraction is slow!

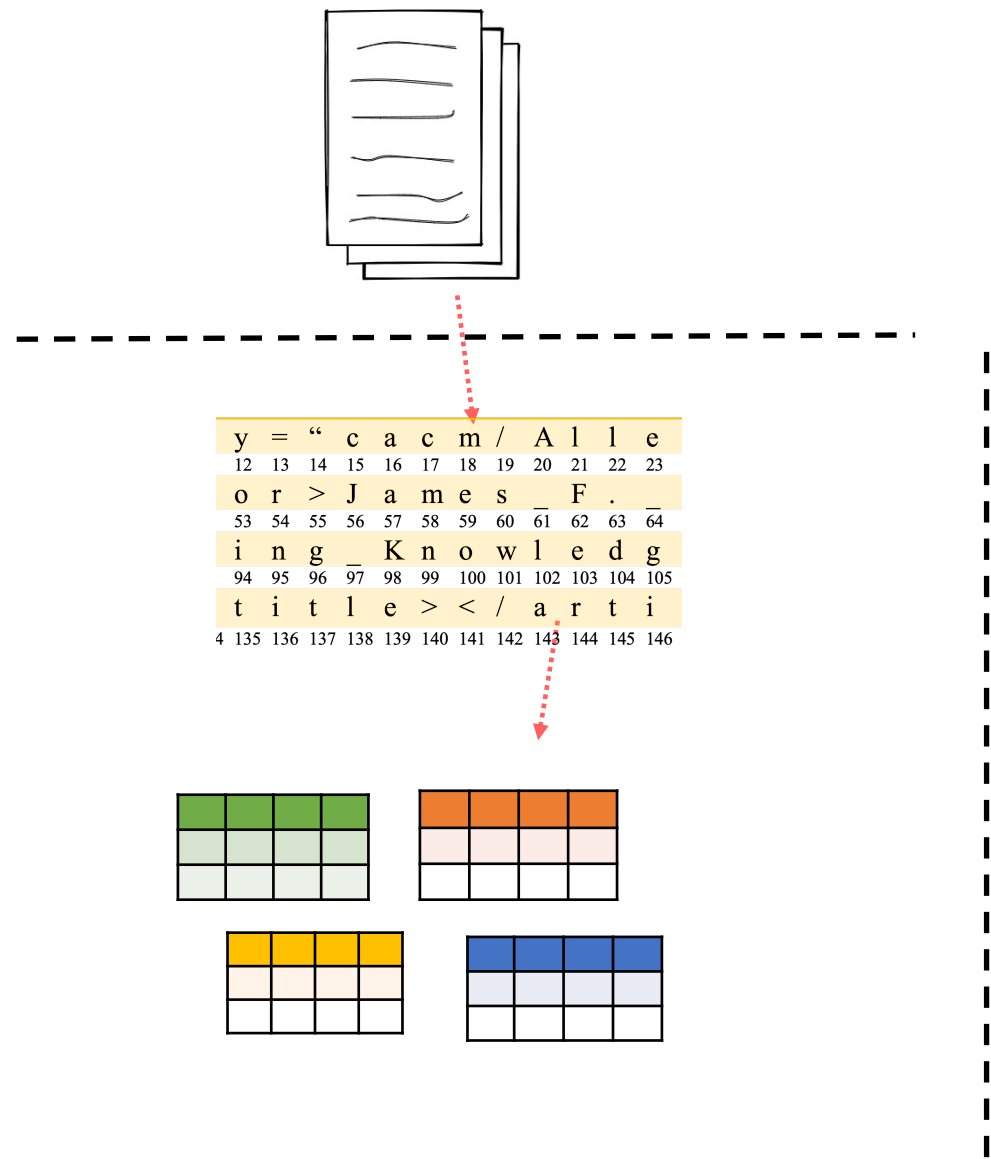
Worse

Research Statement

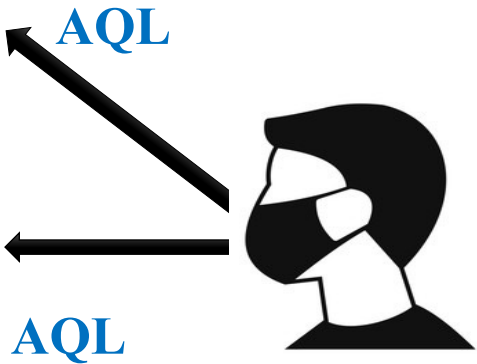
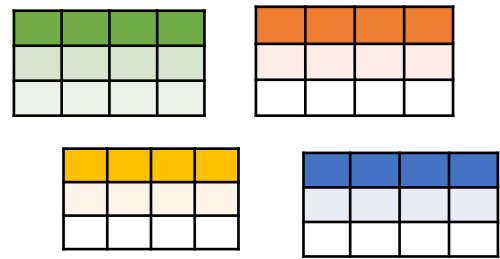
By treating extracted relations as **materialized views** we can design **optimization mechanisms** for information extraction process when **frequent updates** happen to source documents in a **large corpus**.

Realization Setting

- SystemT
- AQL
- Document Spanners

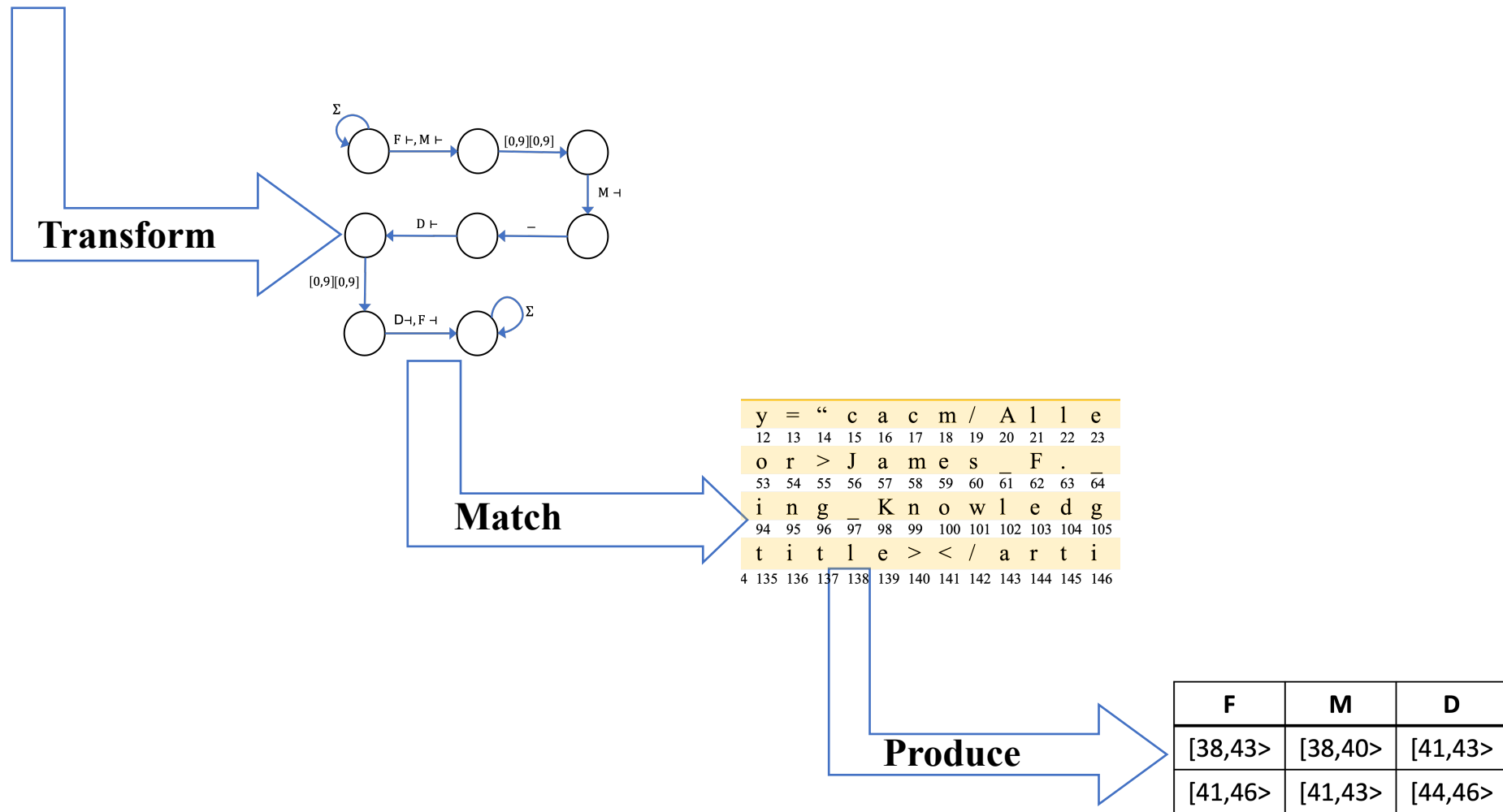


y	=	"	c	a	c	m	/	A	l	l	e
12	13	14	15	16	17	18	19	20	21	22	23
o	r	>	J	a	m	e	s		F	.	
53	54	55	56	57	58	59	60	61	62	63	64
i	n	g		K	n	o	w	l	e	d	g
94	95	96	97	98	99	100	101	102	103	104	105
t	i	t	l	e	>	<	/	a	r	t	i
4	135	136	137	138	139	140	141	142	143	144	145



Matching Model

$$\Sigma^* F\{M\{[0,9][0,9]\} - D\{[0,9][0,9]\}\}\Sigma^*$$



Relational Operators

Join Semantics

$$\Sigma^* F\{[0,9][0,9] - D\{[0,9][0,9]\}\Sigma^* \bowtie \Sigma^* F\{M\{[0,9][0,9]\} - [0,9][0,9]\}\Sigma^*$$

F	D
[38,43>	[41,43>
[41,46>	[44,46>



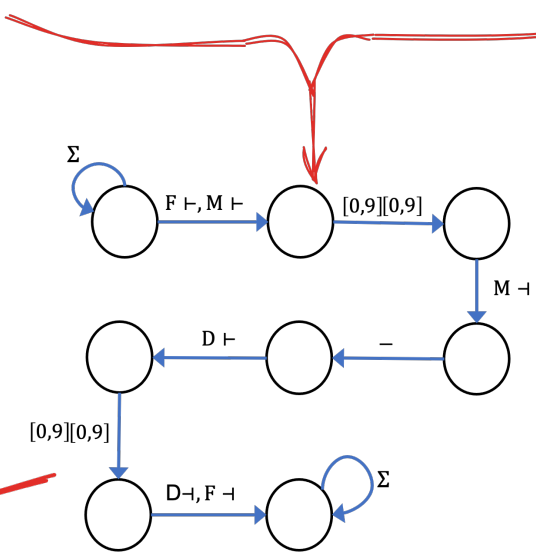
F	M
[38,43>	[38,40>
[41,46>	[41,43>

==

F	M	D
[38,43>	[38,40>	[41,43>
[41,46>	[41,43>	[44,46>

Efficient Construction of Spanners*

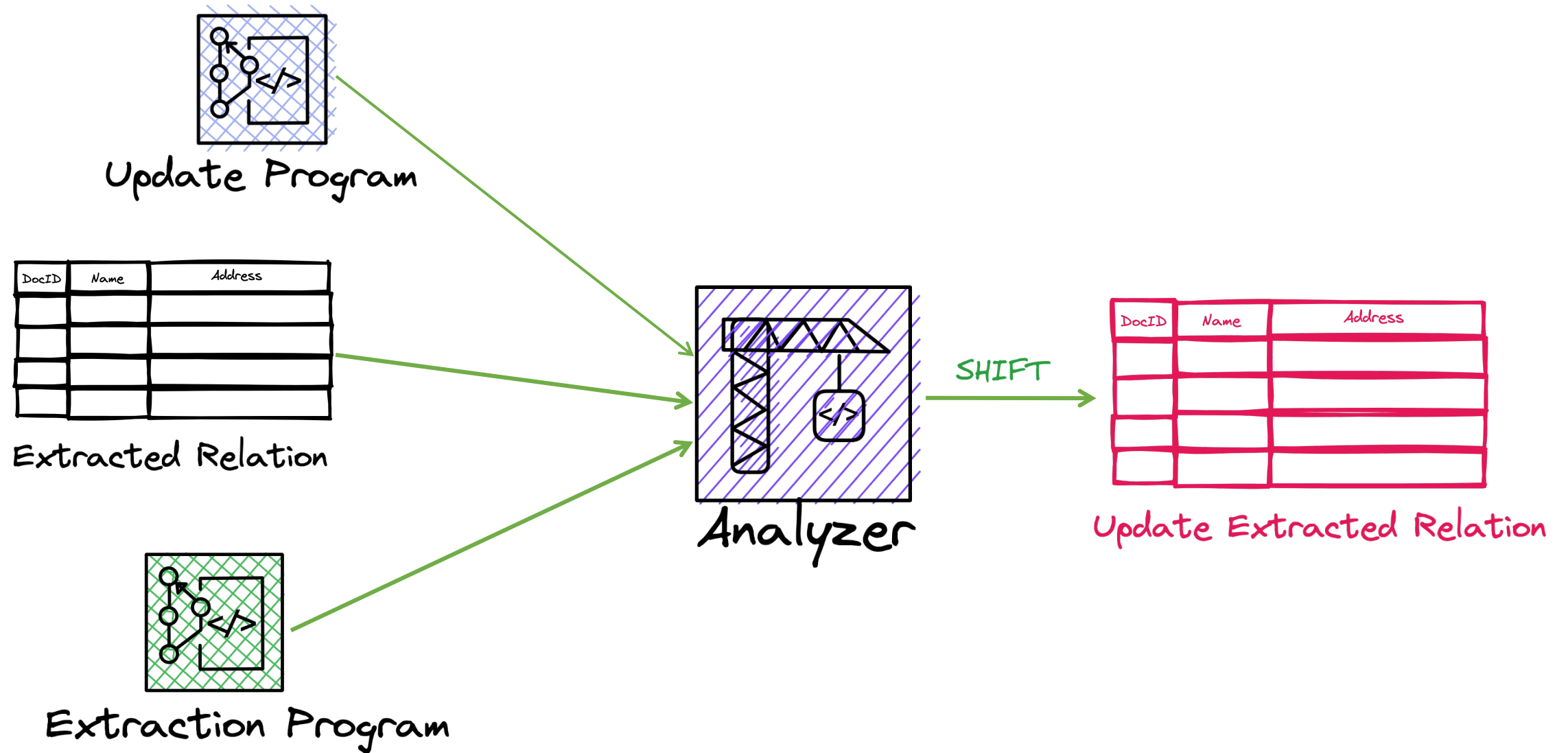
$$\Sigma^* F\{[0,9][0,9] - D\{[0,9][0,9]\}\Sigma^* \bowtie \Sigma^* F\{M\{[0,9][0,9]\} - [0,9][0,9]\}\Sigma^*$$



Update Categories



Approach Summary



On **2023/24/03**, we rented and watched 'A Man
 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41
 an'. I highly recommend it as an inspiring
 42 43 44 45 46 47 48 49 50 51 52 53 54 55 56 57 58 59 60 61 62 63 64 65 66 67 68 69 70 71 72 73 74 75 76 77 78 79 80 81 82
 g and humorous film to enjoy.
 83 84 85 86 87 88 89 90 91 92 93 94 95 96 97 98 99 100 101 102 103 104 105 106 107 108 109 110 111

Original Relation

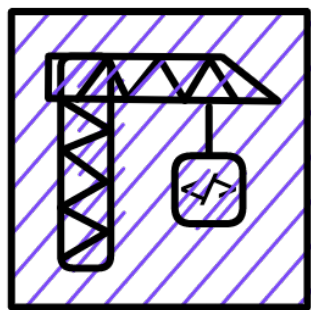
A	T
[70, 79>	[93, 97>
[84, 92>	[93, 97>



Updated Relation

A	T
[75, 84>	[98, 102>
[89, 97>	[98, 102>

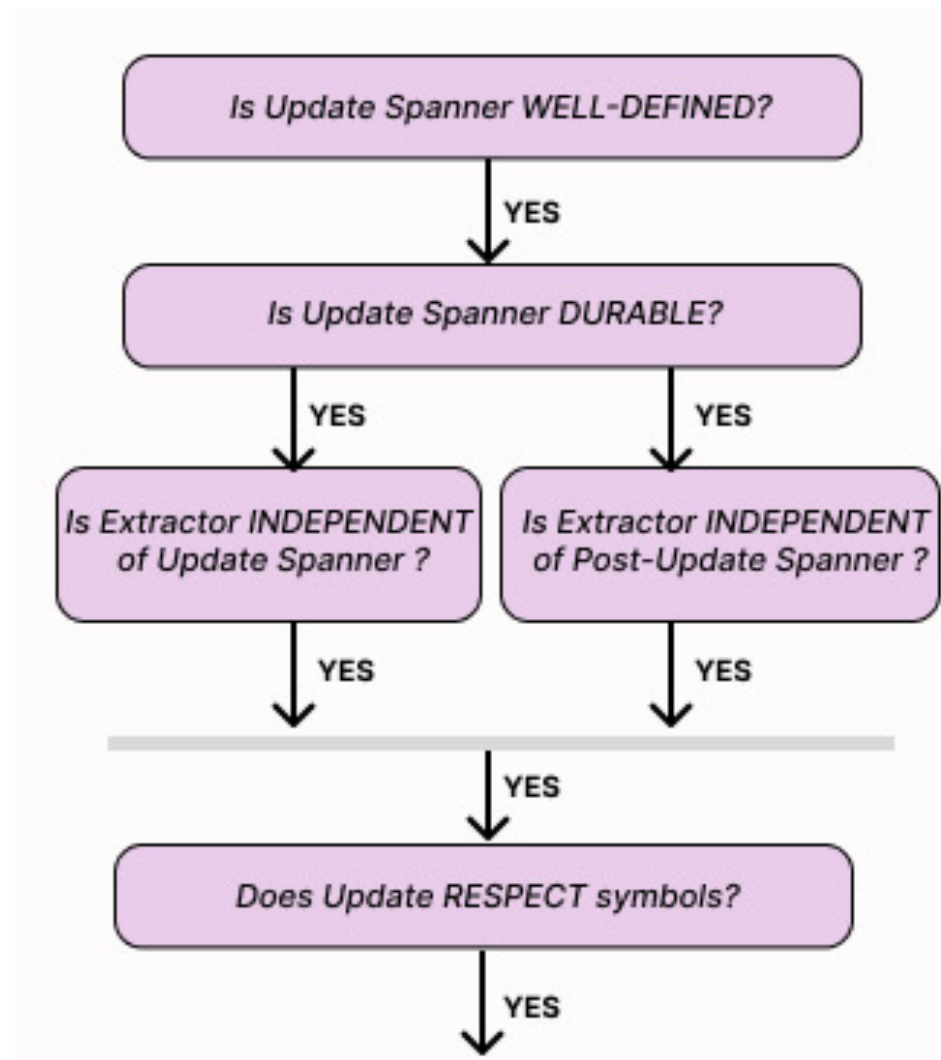
Identifying Shiftable Updates



Devised Tool

Performs

**Polynomial
Time
Complexity**



shiftable

Practicality of Detecting Shiftable Updates



Ext.	Des.	Est. Num. of Matched Doc.	Est. Num. of Ext. Records
SJrn	article title, pages, year, volume, and journal from journal articles	1,422,100	10,887,480
SMod	titles of publications last modified in 2015	17,820	47,480
SAA+	first authors of pubs with 2+ authors	2,304,220	23,011,440
S2010	titles of publications from 2010	302,380	1,264,360
SVLDB	first authors of publications from VLDB	1,860	5,360
SVAA+	multi-author VLDB articles	1,680	4,920
SJ2020	journal titles from 2010	132,360	380,080
S2010Δ	titles from 2010 or updated in 2015	Time out	Time out

Prim.

Comp.

Update	Des.	Est. Num. of Updated Doc.	Est. Num. of Updated Records
Classifying Hashtags	Annotate hashtags	20	20
Enabling URLs.	Update URLs to become clickable links	881,740	6,711,640
Reversing Month and Day in Dates	Convert from yyyy-mm-dd to yyyy-dd-mm	3,091,200	34,870,160
Changing DOIs	Update to convert one format to the other	2,830,180	22,425,760

Corpus: [DBLP](#) Number of documents: [3,091,270](#) Corpus size: [7.9 GB](#) Threaded: [Single threaded](#) Machine: [Ubuntu 20.04.1 LTS \(Focal Fossa\)](#) and [AMD EPYC 7502P 32-Core Processor](#).

Independent of the Database size

Faster than Re-extraction

min	avg	max
8,400 ± 725	50,943	156,400 ± 26042*
*of those < 450,000		

Extraction Time (sec)

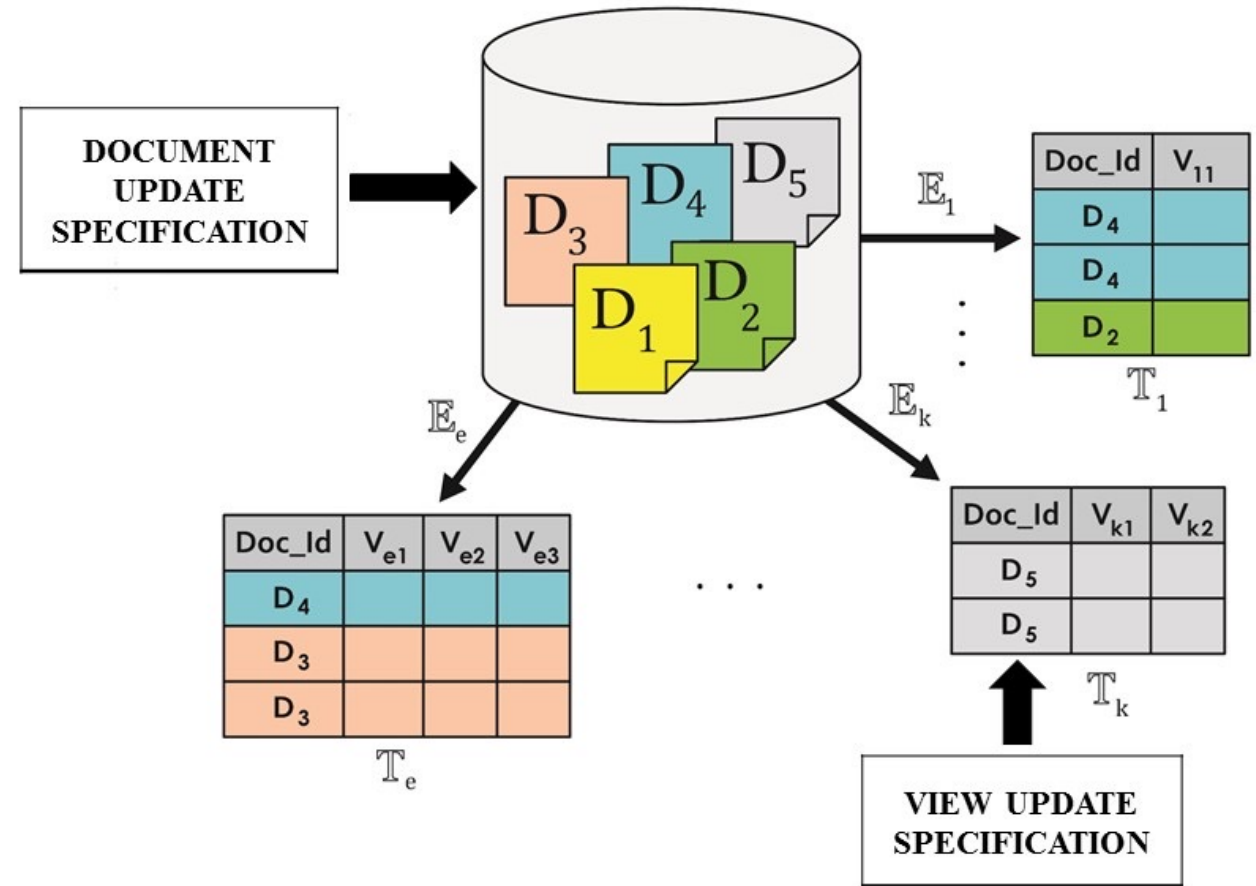
Update	Verification			Shift		
	min	avg	max	min	avg	max
Classifying Hashtags	20	64	127	0.2±0.4	10.2	39.2±76.8
Enabling URLs.	36	108	212	1.1±0.1	67.1	239.3±1.8
Reversing Dates	1	33.6	64	1.5±0.2	78.7	273.5±6.5
Changing DOIs	18	58.2	110	1.2±0.2	72.4	252.4±7.6

Maintenance Time (sec)

Corpus: DBLP Number of documents: 3,091,270 Corpus size: 7.9 GB Threaded: Single threaded Machine: Ubuntu 20.04.1 LTS (Focal Fossa) and AMD EPYC 7502P 32-Core Processor.

Exposed Research Directions

- More complex update models
- Necessary conditions
- Alternative extraction settings
- Unstructured Data Cleansing
- Graph as information source



Thank You!
Questions?