Near Identity Relationships

No Author Given

No Institute Given

0.1 Near identity relationship

In order to evaluate if EMCs can help to identify pair of entities engaged in a near-identity relationship we provide the definition of signature and the notion of corpus. The definition of signature is related to the definition of the most frequent context in a lattice.

Definition 1 (Mot Frequent Context). *Given an identity (resp. difference, incompleteness) lattice, and considering the order of relations between contexts, the most frequent context is the (no empty) context which accumulates the highest number of pairs.*

Definition 2 (Signature). *Given a list of EMCs, their identity, difference and incompleteness lattices, the signature is the triplet composed with the most frequent context of identity lattice, the most frequent context of difference lattice and the most frequent context of incompleteness lattice.*

In the sequel we use the term *corpus* and the notation c to design a list of EMCs and the notation *sign* for a signature. It is possible to discover several signatures $S = \{sign1, sign2, ...\}$ for a given corpus.

This discovery is done iteratively: once the first signature $sign_1$ has been detected on a corpus c_1, the EMCs that matched $sign1$ are removed from c_1 leading to the creation of a corpus c_2 . We then run a second iteration on c_2 in order to detect the next signature $sign_2$. The iteration stops when a *coverage* reach a given threshold.

For a set of signatures S and given initial corpus c_1, the coverage is defined by the number of EMCs that matches exactly one of the signatures in S over the size of the initial corpus c_1.

$$c(S) = \frac{nb\_match\_EMCs}{size(c_1)}$$

Just as key discovery is used to highlight strong identity relationships, we would like to highlight weak identity relationships using signatures. Signatures discovery is performed on a corpus and consists in searching the most frequent contexts in this corpus. The corpus provided must be representative of a weak identity relationship: for example individuals have been linked because they refer to the same general concept such as, the book's of a writer's work or films of a director's filmography. In this use case we want to evaluate if a weak identity relationship can be described with few signatures. In other words, obtaining a coverage $c(S)$ over 80% with a limited set of signatures S.

Corpora construction We have built 5 corpora representing 2 different categories of near entity relationship. (i) relations that describe a more general concept than the one encoded in knowledge graphs (the concept of a literary or cinematographic work versus the concept of a book or film) and (ii) relationships that describe much more tenuous links between entities, entities linked together by their country. What motivated the construction of the second category was the construction of entity spaces as described by [Van Erp et *al.* Toward Entity Spaces] , where the label Germany appears in three different contexts: the context of the meat industry, the context of the German population and the context of the German Davis Cup team. The table 1 presents for each corpus the type of entities used to build the pair, the property used to link these entities, and the number of author, director or countries present in each corpus.

Corpus name	Entity 1 type	Entity 2 type	Link done on property	Nb
Literary Work: books written by the same author	DBpedia Book	YAGO Book	author (created-inv)	4928 authors
Film Work: films made by the same director	DBpedia Film	YAGO Film	director (created-inv)	8500 film directors
Book University: books and universities located in the same country	DBpedia Book	DBpedia University	islocatedin	126 countries
Book Mountain: books and mountains located in the same country	DBpedia Book	DBpedia Mountain	islocatedin	143 countries
Mountain University: mountains and universities located in the same country	DBpedia Mountain	DBedia University	islocatedin	598 countries

Table 1. Description of the 5 corpora.

Signature Detection To explicit signatures detection we present here a toy example. The table 2 presents a corpus computed from pairs of books of Agatha Christie. The pairs share the same author in each case, but titles (label) differs and the number of pages or the isbn number are missing. The most frequent identity context is $\{created - inv\}$, the most frequent difference context is $\{label\}$,

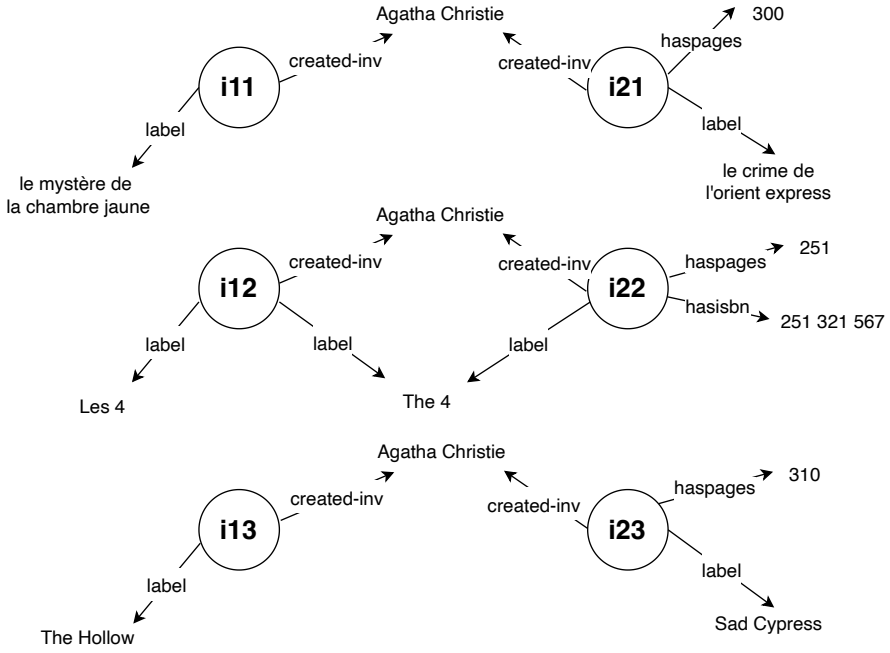
the most frequent incompleteness context is $\{haspages\}$. The last line of the table represents the signature as the concatenation of the 3 most frequent contexts. Notice that in this toy example, all the EMCs matches with the first signature detected (we obtain a coverage of 100%) we do not need a second iteration.

Signature Discovery and Coverage In the same idea we performed signature discovery (i) on the Literary Work and Film Work corpora and (ii) on the 3 corpora of books, mountains and universities from the same country. The first row of Table 3 shows that 90% of the pairs from the literary works corpus are recognized by a single signature $s1 = \{created - inv\}, \{skos : preflabel\}, \{wascreatedonyear\}$. The second line shows that 2 signatures ($s2$ and $s3$) need to be combined for 90% of the pairs in the cinematographic corpus to be recognized. The second part of Table 3 shows that 96% of pairs from the book-mountain and book-university corpora are recognized by the same signature. But it takes a combination of 4 signatures to recognize 98% pairs of the mountain-university corpus. It appears that, on the 5 corpora studied, it is possible to summarize near-identity relationships with few singatures.

0.2 Code

Corpora construction, signature detection and coverage computation are available in the following scripts:

- iswc2024/pattern/author_work_pattern.py for the corpus author work
- iswc2024/pattern/director_work_pattern.py for the corpus cinematographic work
- iswc2024/pattern/country_work_pattern.py for the 3 corpora of entities linked by their countries



i11,i21	{created-inv}	{label}	{haspages}
i12,i22	{created-inv,label}	{label}	{haspages,hasisbn}
i13,i23	{created-inv}	{label}	{has-pages}
signature	{created-inv}	{label}	{has-pages}

Table 2. The example of signature construction based on 3 EMCs.

corpus	signature $\varepsilon\Delta\Omega$	nb match	total nb EMCs	coverage
literary work	s1={created-inv},{skos:preflabel},{wascreatedonyear}			
	s1	148281	165508	0.90
film work	s2={directed-inv}{skos:preflabel},{wascreatedonyear} s3={directed-inv}{skos:preflabel},{islocatedin}			
	s2 ∪ s3	606529	674952	0.90
book mountain	s4={islocatedin},{skos:preflabel}, {created-inv}			
	s4	15320925	15993180	0.96
book university	P5={islocatedin},{skos:preflabel},{created-inv}			
	P5	15320840	15993082	0.96
mountain university	s6={islocatedin},{islocatedin},{haslatitude}, {haslongitude} s7={islocatedin},{islocatedin},{graduatedfrom-inv} s8={islocatedin}{islocatedin},{} s9={islocatedin}{islocatedin},{hasmotto}			
	s6 ∪ s7 ∪ s8 ∪ s9	4708646	4795678	0.98

Table 3. signatures detection on corpora.The coverage is computed as follows: $\frac{nb\_EMCs\_that\_matches\_pattern}{nb\_total\_EMCs}$