Investigating the quality of a bibliographic knowledge base using partitioning semantics

Léa Guizol, Madalina Croitoru

Investigating the quality of a bibliographic knowledge base using partitioning semantics

Léa Guizol and Madalina Croitoru

***Abstract*—With the aim of evaluating and improving link quality in bibliographical knowledge bases, we develop a decision support system based on partitioning semantics. Two such semantics have been proposed, the novelty of this approach consisting on using symbolic values criteria for partitioning. In this paper we investigate the limits of those partitioning semantics: how the characteristics of the input (objects and criteria) influences characteristics of the result, namely correctness of the result and execution time.**

I. Real word application

The Sudoc (the catalogue of French university libraries) is a bibliographic knowledge base which contains bibliographic notices (document descriptions $\approx 10.000.000$), and authority notices (person descriptions $\approx$ 2.000.000). An authority notice possesses some attributes (ppn[1], names set, date of birth...). A bibliographic notice also possesses some attributes (title, ppn[1], language, publication date, domains codes list) and link(s) to authority notices. A link is labeled by a role (as *author*, *illustrator* or *thesis advisor*) and means that the person described by the authority notice has participated as the labeled role to the document described by the bibliographic notice.

Example 1 (Link): Sophocle is represented by the authority notice with 027143619 ppn. His known names according to the authority notice are "Sophocle", "Sofokles", "Sofocle", and "Sophocles Atheniensis tragicus". His date of birth according to the authority notice is "0496? av. J.-C.".

Sophocle has contributed as author to the theatre play *Antigone* represented by the bibliographic notice with 143334670 ppn, which has "Antigone" as title and "2009" as publication date.

So, there is a link labelled "author" between bibliographic notice 166858013 and authority notice 027143619.

A typical input of a book, by a librarian, in Sudoc takes place as follows. The librarian enters the title of the book, ISBN, number of pages and so forth (referred later on as the attributes of the bibliographic record). Then (s)he needs to indicate the authors of the book. This is done by searching in the Sudoc base for the person authority reference of that name. If several possibilities are available (e.g. homonyms), the librarian decides based on the bibliographic information associated to each candidate which one is most suitable for the choice of author of the book at hand. So, (s)he looks existing links in Sudoc. If none of the authors in the base is suitable, then the librarian will create a new authority record in the system and link the book to this new record. The lack of distinguishing characteristics in the authority records and the lack of knowledge about the identity of the book's author imply that the librarian's decision is based on consultation of previous bibliographic records linked to each considered candidates. So any linkage error will entail new linkage errors.

Example 2 (An erroneous link in Sudoc): The authority notice of ppn 082030936 represents a "Bessière, Christian" person which illustrates children books. The authority notice of ppn 070947155 represents a "Bessière, Christian" person which is a computer science researcher. The bibliographic notice with 123728525 ppn represents a book with "Principles and Practice of Constraint Programming" as title and written by a "Bessière, Christian".

Unfortunately, this bibliographic notice is linked to the authority notice of ppn 082030936, which represents the children book illustrator, and not the computer science researcher. This is an erroneous link that might persuade librarians to link to the children book illustrator new bibliographic notices representing books about computer science, creating new linkage errors.

In this paper we consider the methodology (presented in [11] and resumed in section II) to assess the correctness of Sudoc links. This methodology is based on two Entity Resolution methods (presented in [14] and resumed in section III). These Entity Resolution methods are partitioning semantics. ***The contribution of this paper is the investigation of the limits of those partitioning semantics: how the characteristics of the input (criteria and objects) influences characteristics of result (correctness of result and execution time to get it)***.

Let us start by presenting the global methodology of our partitioning approach.

II. Basic notions

We choose to see Sudoc bibliographic notices from one of their contributor's view point called **contextual authority** (this contributor is the C contributor, and the i contextual authority is denoted Nc_i). A contextual authority represents a contributor link from Sudoc. A contextual authority possesses all attributes of its bibliographic notice (title, publication language, publication date, publication domain), and 3 other attributes depending on the C contributor:

Léa Guizol and Madalina Croitoru working at LIRMM (University of Montpellier II & CNRS) and INRIA Sophia-Antipolis, France. (email: {guizol, croitoru}@lirmm.fr).

This work has been supported by the Agence Nationale de la Recherche (grant ANR-12-CORD-0012). We are thankful to Mickaël Nguyen for his support.

[1]A ppn identifies a notice.

- appellation: the appellation list of the authority notice which represents the contributor C;
- role of the C contributor;
- other contributors: list of the authority notices which represent all contributors of the bibliographic notice but not the C contributor.

Example 3 (Attributes of a contextual authority): Let us take the link between the authority notice 027143619 and the bibliographic notice 143334670 presented in example 1. We construct the contextual authority which represents this link. Its attributes are:

- title: "Antigone"
- publication date: "2009"
- publication language: "fre" (means "french")
- publication domains list: {} (there is no given publication domains for this contextual authority)
- appellation: {"Sophocle", "Sofokles", "Sofocle", and "Sophocles Atheniensis tragicus"}
- role of C contributor: "author"
- other contributors: $\{032756534\}$ (032756534 represents the "Brunet, Philippe" which has translated the theatre play in French.)

In the following, we will denote:

- $C(Nc_i)$, the C contributor of the contextual authority Nc_i;
- $Nb(Nc_i)$, the bibliographic notice from which Nc_i has been created;
- $Na(Nc_i)$, the authority notice from which Nc_i has been created (corresponds to the authority notice which represents $C(Nc_i)$).

To investigate the quality of Sudoc links comes down to the following steps:

1) Choose an appellation A (as "Harris, Sam").
2) Select every authority notice having an appellation close to the appellation A.
3) Construct all contextual authorities corresponding to a link between any Sudoc bibliographic notice and an authority notice selected in step 2.
4) Group contextual authorities such as contextual authorities are together if and only if they are close together .
5) Question links if a contextual authority Nc_i is not in a class containing all and only the contextual authorities Nc_j such as $Na(Nc_i) = Na(Nc_j)$.

The step 2 is a way to divide the problem of detecting erroneous links in Sudoc into detecting erroneous links in smaller Sudoc subsets, easier to manage (similar to blocking [17], canopies [19] methods, etc). The set of contextual authorities of step 3 is **the Sudoc subset related to appellation A**. Later we will be interested in analysing such contextual authorities subsets. However, **in this paper, we will focus on step 4 which is solved as an entity-resolution problem**. Let us talk about the entity resolution approach used to detect erroneous links in Sudoc.

III. Entity resolution approach for Sudoc

To determine whether contextual authorities must be "together" or not, we partition[2] them. For a given partition, two contextual authorities are "together" if they are in a same class, and "separated" if not.

We will summarize in this section the clustering semantics presented in [14] (please consult [14] for further details): what is a best partition (in general case section III-C, according to global and local semantics in sections III-D and III-E). The used algorithms are explained in section III-F. We will define two special partitions in the subsection III-B useful to answer to the link quality issue for the Sudoc knowledge base. To partition a set of contextual authority, we use symbolic criteria. Let us define them.

A. Symbolic criteria

A symbolic criterion is a function which gives a symbolic comparison value for two objects. This comparison value could be a closeness, farness, $neutral$, $always$ (obligation to put together the objects), or $never$ (obligation to separate the objects) value. For a given criterion, its closeness (respectively farness) value set could contain several values more or less strong and intense. Intuitively, we denote closeness values (respectively farness values) by more or less $+$ (respectively $-$) symbols and the more the value has symbols, the stronger or more intense it is. The $always$ (respectively $never$) value is stronger than any closeness (respectively farness) value. Any closeness (respectively farness) value is stronger than the $neutral$ value. Let c be a criterion, and o_i, o_j be two objects to compare. We denote $c(o_i, o_j)$ the comparison value between o_i and o_j according to c.

Example 4 (Comparison value for a criterion): Let us describe the $date$ criterion. For 100 (respectively 60) years at least between publication dates of two contextual authorities Nc_i and Nc_j to compare, it gives a $--$ (respectively $-$) comparison value. If there is less than 60 years between them, or an unknown date, $date(Nc_i, Nc_j) = neutral$. $--$ is a $date$ comparison value more intense than the $date$ comparison value $-$.

B. Sudoc partitions

Definition 1 (The initial partition, denoted Pi): is the only partition deduced from Sudoc such as two contextual authorities Nc_i and Nc_j are in a same class if and only if $Na(Nc_i) = Na(Nc_j)$: they had the same contextual authority representing their C contributor when they were created.

Example 5 (Initial partition): Let us represent an object set $\mathbb{O}_{hs} = \{Nc_1, Nc_2, Nc_3, Nc_4, Nc_5, Nc_6\}$ in Table 1. Each object is a contextual authority, representing a Sudoc link between a bibliographic notice and an authority notice. Id is the contextual authority identity. For each Nc_i of them, the appellation is the $C(Nc_i)$ appellation and the ppn is a way to identify $Na(Nc_i)$.

[2]A **partition** P of an object set X is a set of **classes**(subsets of X) such as each object of X is in one and only one class of P.

id	title	date	domain	appellation	ppn
Nc_1	Letter to a Christian nation		religion	"Harris, Sam"	1
Nc_2	Surat terbuka untuk bangsa kristen	2008	religion	"Harris, Sam"	1
Nc_3	The philosophical basis of theism	1883	religion	"Harris, Sam"	1
Nc_4	Building pathology	2001	building	"Harris, Samuel"	2
Nc_5	Building pathology	1936	building	"Harris, Samuel"	2
Nc_6	Aluminium alloys 2002	2002	physics	"Harris, Samuel"	2

TABLE I
EXAMPLE OF CONTEXTUAL AUTHORITIES

$C(Nc_1)$ and $C(Nc_2)$ represent a same person, as $C(Nc_4)$, $C(Nc_5)$ does. The initial partition on $\mathbb{O}_{hs}$ is the partition which puts contextual authorities in the same class if and only if their C contributor ppn is the same. It is: $Pi_{hs} = \{\{Nc_1, Nc_2, Nc_3\}, \{Nc_4, Nc_5, Nc_6\}\}$.

Definition 2 (The human partition, denoted Ph): is the perfect partition according to a human expert: two contextual authorities are in a same class if and only if the human expert believes that their C contributor corresponds to a unique real person in the real word. If the human partition Ph is not the same partition as Pi, that means that there is a link problem in Sudoc according to the expert which made Ph.

Example 6 (Human partition): Let us give the human partition (determined by an expert) on the object set $\mathbb{O}_{hs}$ presented in example 5 and in Table 1. This partition is: $Ph_{hs} = \{\{Nc_1, Nc_2\}, \{Nc_3\}, \{Nc_4, Nc_5\}, \{Nc_6\}\}$ because $C(Nc_1)$ and $C(Nc_2)$ represent a same real-word person, as $C(Nc_4)$ and $C(Nc_5)$ do.

Work hypothesis: *we suppose that the human partition Ph is a best partition, and that the initial partition Pi is a best partition if and only if Ph is Pi.*

If the work hypothesis is true, detecting link issues in a Sudoc subset boils down to:

- construct contextual authorities of the Sudoc subset,
- evaluate the initial partition on those contextual authorities (please see the initial partition of a Sudoc subset in example 5),
- evaluate best partitions on those contextual authorities.

The Sudoc subset has a link issue if and only if the initial partition value is not in the best partition values. In order to see if the initial partition is a best partition, let us determine what is a best partition.

C. Partition evaluation: interest, meaning, partitionning semantics

In this part, we will evaluate the partitions on an object set $\mathbb{O}$ according to a symbolic criteria set $\mathbb{C}$ and order partitions by their values.

1) Valid partitions: Firstly, we are only interested in valid partitions [2]. A partition P is valid if and only if there are no two objects o_i,$o_j \in \mathbb{O}$ such as:

- they are in a same class of P and they $never$ have to be together according to a $\mathbb{C}$ criterion (expressed by the $never$ comparison value), or
- they are in distinct P classes but should $always$ be together according to at least a $\mathbb{C}$ criterion.

Example 7 (Valid partition): Human and initial partitions (respectively Ph_{hs} and Pi_{hs}) given on examples 6 and 5 on $\mathbb{O}_{hs}$ are valid according to the criteria set $\mathbb{C}_{hs} = \{title, domain, date\}$. Indeed, the only criteria which can give an $always$ or $never$ value in $\mathbb{C}_{hs}$ is $title$, which gives the $always$ value only when comparing two contextual authorities with an identical title. So, $title(Nc_4, Nc_5) = always$ and Nc_4, Nc_5 are in a same class for both partitions so Ph_{hs} and Pi_{hs} are valid.

However, the following partition is not valid according to $\mathbb{C}_{hs}$ because Nc_4 and Nc_5 are not in a same class: $P'_{hs} = \{\{Nc_1, Nc_2\}, \{Nc_3\}, \{Nc_4\}, \{Nc_5\}, \{Nc_6\}\}$.

2) Intra and inter classes values: Secondly, a partition P has an intra value and an extra value per criterion. The intra value of a criterion c depends on the most intense farness (please see section III-A) value of c such as it compares two objects in a same class (should not be the case according to c). In the same way, the inter value of c depends on the most intense closeness value of c such as it compares two objects in distinct P classes. The inter value measures proximity between classes and the intra value measures distance between objects in a class[13]. We note that the $neutral$ comparison value does not influence partition values.

Example 8 ($neutral$ value influence): Let $\mathbb{O} = \{o_1, o_2, o_3\}$ be an object set and ex be a criterion which compares objects of $\mathbb{O}$. We put $ex(o_1, o_2) = ++$, $ex(o_1, o_3) = neutral$ and $ex(o_2, o_3) = neutral$.

$ex(o_1, o_2) = ++$, so ex prefers partitions such as o_1 and o_2 are in a same class. That implies that $\{\{o_1, o_2, o_3\}\}$ has a better value as $\{\{o_1\}, \{o_2, o_3\}\}$. However, $\{\{o_1, o_2, o_3\}\}$ has the same value than $\{\{o_1, o_2\}, \{o_3\}\}$ because $neutral$ does not change anything to partition values and $ex(o_1, o_3) = ex(o_2, o_3) = neutral$.

3) Best partition: A partition P is a best partition according to a criteria set $\mathbb{C}$ if and only if P has a best partition value according to a criteria set $\mathbb{C}$.

A partition P has a best partition value according to a criteria set $\mathbb{C}$ if P is valid and it is impossible to improve an inter or intra value of one $\mathbb{C}$ criterion without decreasing an inter or intra value of at least a $\mathbb{C}$ criterion (Pareto equilibrium [22]). Let us see how this definition is used for global and local semantics.

Example 9 (Best partition): Human and initial partitions (respectively Ph_{hs} and Pi_{hs}) given on examples 6 and 5 on $\mathbb{O}_{hs}$ are valid according to the criteria set $\mathbb{C}_{hs} = \{title, domain, date\}$ (as explained in example 7). We consider that:

- $title$ criterion gives an $always$ value if it compares two contextual authorities with the same title;
- $domain$ criterion gives a $+$ (closenness) value if it compares two contextual authorities with the same domain, and a $-$ (farness) value if domains are distinct;
- $date$ criterion gives a $-$ value if it compares two contextual authorities with publication dates 60 to 100 years distant from each other, and a $--$ value if they are more than 100 years distant;
- each criteria gives the $neutral$ value in other cases, especially when a compared attribute is missing.

The initial partition Pi_{hs} is not a best one because Nc_5 and Nc_6 have dates more than 60 years distant ($date(Nc_5, Nc_6) = -$) and distinct domains ($domain(Nc_5, Nc_6) = -$) but are in a same class. To put Nc_6 in a new distinct class will not improve the $date$ intra value (because $domain(Nc_2, Nc_3) = --$ and Nc_2, Nc_3 are also in a same class) but it will improve the $domain$ intra value without decreasing any criterion intra or extra values.

We saw in a general case how to give a value to a partition, and how to determine whether a partition has a best value. Let us explain how this is used by global and local partitioning semantics.

D. Global semantics

The global semantics uses the best partition definition (defined section III-C3) for an object subset $\mathbb{O}$ (a partition on an object set has a single partition value). However, separately or wholly partitioning two objects sets which have nothing in common yields different results as seen in [14].

Example 10 (Global semantics evaluating a partition on an object set $\mathbb{O}$): Let us consider the human partition Ph_{hs} on the object set $\mathbb{O}_{hs}$ given in example 6 and the criteria set $\mathbb{C}_{hs}$ given on example 9. This partition, according to global semantics and with respect to the criteria set $\mathbb{C}_{hs}$ is not coherent with some of $\mathbb{C}_{hs}$ criteria. The P_{hs} value is such that:

- inter classes $domain$ value is very bad ($always$) because Nc_1 and Nc_2 are in distinct classes but are both about religion.
- intra classes $date$ value is bad ($--$) because Nc_4 and Nc_5 are in a same class, but with between 100 and 60 years distant publication dates.

However, P_{hs} has a best partition value because increasing an inter or intra value (as inter $domain$ value by merging $\{Nc_1, Nc_2\}$ and $\{Nc_3\}$ classes) is not possible without decreasing another (Nc_2 and Nc_3 have publication dates more than 100 years distant, so putting them in a same class will decrease the $date$ intra value).

E. Local semantics

The local semantics, when evaluating a partition on an object set $\mathbb{O}$ with respect to a criteria set $\mathbb{C}$ gives a partition value per incoherent part of $\mathbb{O}$ and for the coherent part of $\mathbb{O}$. An **independent part** $\mathbb{O}_a$ is a subset of $\mathbb{O}$ such as:

- there is no $c(o_i, o_j)$, an $always$ or closeness value with $o_i \in \mathbb{O} - \mathbb{O}_a$, $o_j \in \mathbb{O}_a$, and $c \in \mathbb{C}$;
- there is no subset of $\mathbb{O}_a$ for which the previous property is true;

An **incoherent part** $\mathbb{O}_d$ is an independent part such that there is also $b(o_k, o_l)$, a farness or $never$ value such that $o_k, o_l \in \mathbb{O}_d$ and $b \in \mathbb{C}$.

The partition value of an incoherent part is based on each comparison between objects which are in it. The **coherent part** partition value of $\mathbb{O}$ is based on each comparison between objects which are not in a same incoherent part of $\mathbb{O}$.

Example 11 (Independent and incoherent parts): Let us identify incoherent subsets of the object set $\mathbb{O}_{hs}$ according to $\mathbb{C}_{hs}$ as it has been done in example 10. Let us determine the independent and incoherent parts on $\mathbb{O}_{hs}$.

Nc_1, Nc_2, Nc_3 are close together due to $domain$ criterion: they are about "religion" publication domain. Nc_1, Nc_2, Nc_3 are not close to Nc_4, Nc_5 or Nc_6 according to any of $\mathbb{C}_{hs}$ criteria, so, $\{Nc_1, Nc_2, Nc_3\}$ is an independent part of $\mathbb{O}_{hs}$.

The same way, Nc_4, Nc_5 are close together according to $title$ and $domain$ criteria, but not close to Nc_6: $\{Nc_4, Nc_5\}$ and $\{Nc_6\}$ are independent parts. So, there are three independent parts in $\mathbb{O}_{hs}$: $\{Nc_1, Nc_2, Nc_3\}$, $\{Nc_4, Nc_5\}$ and $\{Nc_6\}$.

Furthermore, $\{Nc_1, Nc_2, Nc_3\}$ and $\{Nc_4, Nc_5\}$, are incoherent parts because they are independent parts and Nc_2, Nc_3 (respectively Nc_4, Nc_5) are far from each other according to the $date$ criterion. Indeed, $date(Nc_2, Nc_3) = --$ (respectively $date(Nc_4, Nc_5) = -$). $\{Nc_6\}$ is not an incoherent part (independent parts containing a single object cannot be incoherent parts).

An incoherent part is an independent part which contains at least an **incoherence**, which is a property of a subset of $\mathbb{O}$ objects such that these objects must be in a same class according to some $\mathbb{C}$ criteria ($title(Nc_4, Nc_5) = always$) and must be in distinct classes according to at least one of $\mathbb{C}$ criteria ($date(Nc_4, Nc_5) = -$). An example is (Nc_4, Nc_5) in example 11.

A partition on $\mathbb{O}$ is a best partition for local semantics if it has best partition values for each incoherent part of $\mathbb{O}$ and for the $\mathbb{O}$ coherent part. Finding all best partition values for an object set is a distributive function for the union of independent objects sets in this semantics. This is due to the fact that the best partition value for coherent parts is always the same (there is no contradiction in coherent part between criteria).

Example 12 (Local semantics evaluating a partition on an object sets $\mathbb{O}$): In example 11, we identified the incoherent parts of the object set $\mathbb{O}_{hs} = \{Nc_1, Nc_2, Nc_3, Nc_4, Nc_5, Nc_6\}$ according to the criteria set $\mathbb{C}_{hs} = \{title, domain, date\}$.

The human partition on $\mathbb{O}_{hs}$ is, as in example 10: $Ph_{hs} = \{\{Nc_1, Nc_2\}, \{Nc_3\}, \{Nc_4, Nc_5\}, \{Nc_6\}\}$. Ph_{hs}, according to local semantics, has 3 partition values, one for the

coherent part and 2 for incoherent parts (1 per incoherent part):

- a perfect value for the coherent part of $\mathbb{O}_{hs}$;
- the incoherent part $\{Nc_1, Nc_2, Nc_3\}$ has a bad inter value for the $domain$ criterion ($always$);
- the incoherent part $\{Nc_4, Nc_53\}$ has an bad intra value for the $date$ criterion ($--$);

This semantics allows us to separate and isolate incoherences in the object set in smaller objects subsets, and to evaluate them separately.

We explained how local and global semantics evaluate partitions. Let us explain how the associated algorithms find all best partition values according to these semantics.

F. Algorithms to find best partition values

Let us briefly explain how algorithms used to find all best partition values for both local and global semantics work (please see [15] for details).

1) The algorithm for global semantics (denoted $globalAlgorithm$): finds all best partition values for an object set $\mathbb{O}$ and a criteria set $\mathbb{C}$ by calculating and evaluating all **reference partitions**[14]. There is a total order between reference partitions such that when one has an **optimal value** (a best intra value for each criterion), and there is no need to test the worse reference partitions (called its **descendants**) to find all best partition values. The reference partitions set is defined for a criteria set. When a reference partition has an optimal value, we do not test its descendants. A better reference partition is tested until there are no more reference partitions to test.

To test a reference partition has a $\mathcal{O}(m \log n)$ complexity for n objects and m comparison values. There are $(k+1)^c$ reference partitions to test in the worst case, with c, number of criteria and k depending on maximum number of criteria closeness comparison values for a $\mathbb{C}$ criterion. The complexity of $globalAlgorithm$ is: $\mathcal{O}((k+1)^c * m \log n)$ in the worst case.

2) The algorithm for local semantics (denoted $localAlgorithm$): finds each incoherent part ($\mathcal{O}(m \log n)$ complexity) and uses $globalAlgorithm$ on each of them. There are $n/2$ incoherent parts at most, so $localAlgorithm$'s complexity is $\mathcal{O}(n * (k+1)^c * m \log n)$ in the worst case.

Let us get to the main paper contribution now and explore how criterion and object characteristics influence results and execution time of the algorithms explained above.

IV. Analysing link quality

In this section, we study ***how input characteristics (data ambiguity, criteria accuracy, number of incoherences) influence the partitioning semantics' output (human partition value, number of incoherences and algorithm execution time)***.

Considered input characteristics are:

- **data ambiguity**: We say that **data** is **ambiguous** if there exists distinct real world entities that have the same name and contributed to some very similar contextual authorities.
- **criteria accuracy**: The criteria can lack precision, and consider as similar distinct contextual authorities (or consider as far close contextual authorities).
- **the number of incoherences** (please see section III-E). As we will see, the number of incoherences is a consequence of data ambiguity and criteria accuracy and influences result characteristics.

Considered output characteristics are:

- **the human partition's value** (please see definition 2).
- **the number of incoherences** (please see section III-E).
- **execution time** of algorithms $globalAlgorithm$ and $localAlgorithm$ (please see section III-F for details and complexity of algorithms).

The currently used criteria are $date$ (time distance between publication dates), $appellation$ (do the C contributors have an appellation which could refer to the same real word person?), $title$, $language$, $domain$ (the list of publication domains of each contextual authority is compared), $otherContributors$ (is there at least a not-C contributor in common?), $role$ (of C contributor), $thesis$ (check if several thesis have been published at the same time), $thesisAdvisor$ (check if a person has been a thesis advisor before finishing his own thesis).

In the following, we explore consequences of each data characteristics on results characteristics. We consider a contextual authorities set $\mathbb{O}$ and a criteria set $\mathbb{C}$.

A. Data ambiguity

Definition 3 (Data ambiguity): The **data ambiguity** is a measure of how many pairs of $\mathbb{O}$ objects there are in an object set $\mathbb{O}$, such that they seem very close (respectively far) but are not in the real life.

Using the above introduced notation, for a set of contextual authorities $\mathbb{O}$, data ambiguity corresponds to how many contextual authorities Nc_i, Nc_j of $\mathbb{O}$ exist such as $C(Nc_i)$ and $C(Nc_j)$ seem to represent the same (respectively distinct) real-word person but do not.

The more data is ambiguous, the more pairs of objects seem close (or far) to each other but are not. **Increasing data ambiguity could increase the number of incoherences** (defined section III-E).

B. Criteria accuracy

Definition 4 (Accuracy of a criterion): The **accuracy of a criterion** c is the ratio of pairs of objects well compared wrt real world (the criterion gives an appropriate closeness, farness, $never$ or $always$ comparison value) to pairs of objects compared with a different than $neutral$ comparison value. For a given criteria set $\mathbb{C}$, the **criteria accuracy** is the average accuracy of $\mathbb{C}$ criteria.

Let us see how criteria accuracy influences results.

The less criteria are accurate, the more pairs of objects are compared with a closeness or $always$ comparison value (respectively farness or $never$ comparison value) according to a criterion but that are far (respectively close) to each other. This could improve the number of incoherences (defined section III-E).

The less criteria are accurate, the more data seems ambiguous, and, the more data is ambiguous, the more it is difficult to make accurate criteria in order to compare the objects.

We saw that decreasing criteria accuracy increase the number of incoherences. Let us see what is influenced by number of incoherences in the next section.

C. Number of incoherences

In an object set $\mathbb{O}$ according to a criteria set $\mathbb{C}$ it is hard to directly measure the number of incoherences (defined section III-E). We will consider in this section that decreasing criteria accuracy increases the number of incoherences (explained in section IV-B), and use it to **observe effects of increasing the number of incoherences** on several characteristics of the results (i.e. human partition value, execution time of $localAlgorithm$ and $globalAlgorithm$, and number of incoherent parts).

1) Human partition having a best partition value: the initial partition (definition 1) is supposed to be a best partition (defined part III-C) if and only if it corresponds to the human partition (definition 2). That implies that the human partition must have a best partition value.

The human partition is supposed to have a best partition value (according to work hypothesis in section III-B). If the initial partition corresponds to the human partition and does not have a best value, that could be because of the number of incoherences as shown in the next example.

Example 13 (Human partition value and number of incoherences): Let $\mathbb{O}_{nm}$ be the contextual authorities Sudoc subset related to "Nicolas Maurice" appellation (as defined in section II). Let $\mathbb{C}$ be the criteria set which contain all criteria described above. According to global semantics, there are 2 best partitions values (the first is good according to all criteria but $language$, the second is good for all criteria, except $title$ and $otherContributors$) and the human partition has a best partition value.

Let us add incoherences by decreasing criteria accuracy. Let $20\%randomDate$ be a new criterion such that, for two contextual authorities with a date attribute in both of them, there is 20% chance to give a random comparison value ($neutral$, a closeness or a farness value). In other cases, it's the same comparison value as the original $date$ criterion ($neutral$ or a farness value). We calculated all best partition values and evaluated the human partition on $\mathbb{O}_{nm}$ for $\mathbb{C}' = (\mathbb{C} \cup 20\%randomDate)$ and found 4 best partitions values. The human partition did not have a best value according to $\mathbb{C}'$.

We can thus see that **increasing the number of incoherences could make the human partition not have a best partition value**. Let us now see how the number of incoherences influences execution time for both algorithms (described section III-F). We start by looking at the execution time for $globalAlgorithm$.

2) Execution time of $globalAlgorithm$: The **execution time for an algorithm** is the time, in milliseconds, took by the algorithm at hand to return a result.

Increasing the **number of incoherences increases the execution time of** $globalAlgorithm$ (explained in section III-D) by requiring more reference partitions to calculate and evaluate with the aim to find out all best partition values (because less reference partitions have an optimal value due to incoherences, which implies that their descendants have to be calculated and evaluated – please see section III-F1).

Example 14 (Number of incoherences and execution time for the $globalAlgorithm$ algorithm): In this example, we took the object set related to appellation "Leroux, Alain" and measure the execution time[3] according to 8 criteria: $appellation$, $title$, $language$, $otherContributors$, $role$, $thesis$, $thesisAdvisor$ and $randomDate'$. The $randomDate'$ is a criterion as $20\%randomDate$ criterion detailed in example 13 but the percentage of random comparison values varies according to "percent of $randomDate'$" of figure 1. For each tested percentage (from 0 to 100% by steps of 5) of random comparison:

- we generated 25 times the $randomDate'$ criterion and measured every time the execution time (represented by a gray dot);
- we calculated the average mark of those 25 execution time measures and represented it by a black dot on the figure.

We can see on figure 1 that the **execution time increases with the percentage of random comparisons** of $randomDate'$ criterion from 1 milliseconds on average for 0% of random comparison (when $randomDate'$ corresponds to $date$ criterion) to 9 milliseconds on average for 100% of random comparison. Also, for a given percentage of random comparisons, the execution time fluctuates only slightly. There is a single spike at 5% of random comparisons, which could be explained by the fact that execution times fluctuates more or 5% of random comparisons than for 10% to 20% of random comparisons.

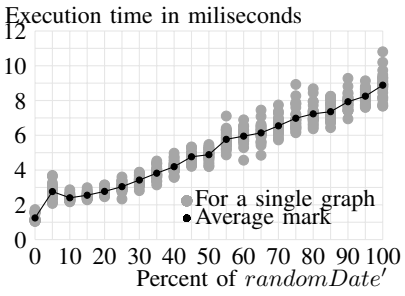

Fig. 1. Execution time to find all best partition values according to global semantics and 8 criteria including $randomDate'$

Therefore, we can conclude that **increasing the number of incoherences increases execution time for** $globalAlgorithm$ in a nearly monotonous manner. Before seeing how the number of incoherences influences $localAlgorithm$'s

[3]We used a Intel(R) Core(TM) i7-2600 CPU 3.40 GHz PC with 4GB of RAM running Windows 7 64 Bit with a Java 1.6 implementation.

execution time, we first need to see **how the number of incoherences influences the number of incoherent parts**.

3) The number of incoherent parts: Incoherences parts are, as explained in section III-E, independent parts which contain at least an incoherence. Therefore, **adding incoherences** can, at the same time, both:

- **increase the number of incoherent parts** (by adding a farness or $never$ value in an independent part not yet incoherent, will add an incoherence in it and make it a new incoherent part);
- **reduce the number of incoherent parts** (by adding a closeness or $always$ comparison value between two objects of two incoherent parts, which will merge them and decrease the total number of incoherent parts of 1).

Example 15 (Incoherences and number of incoherent parts): In this example, we consider the object set corresponding to appellation "Leroux, Alain" and the criteria set of example 14. For each tested percentage $randomDate'$:

- we took the 25 $randomDate'$ criteria generated in example 14 and measured the number of incoherent parts, which is represented on figure 2 by a grey dot;
- we calculated the average mark of those 25 incoherent parts measures and represented it by a black dot on the figure.

For small percentage of random comparison (5% to 15%), the incoherences part number fluctuates a lot, in particular for 5% and when the average number of incoherent parts is more than 1. For no random comparisons or 20% and more random comparisons, there is all the time a single number of incoherent parts.

That example shows that **slightly decreasing the number of incoherences increases the number of incoherent parts** by adding incoherences in independent parts, but **increasing a lot the number of incoherences does not increase number of incoherent parts** because most of the independent parts are merged in a single big incoherent part.

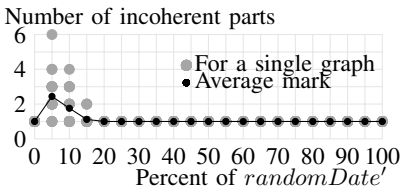

Fig. 2. The number of incoherent parts according to 8 criteria including $randomDate'$

4) Execution time of $localAlgorithm$: for the same reason that the number of incoherences increases the execution time of $globalAlgorithm$, the **number of incoherences also increases the execution time of** $localAlgorithm$.

However, $localAlgorithm$**'s execution time also depends on incoherent parts** because it executes $globalAlgorithm$ for each incoherent part, as shown in section III-E. So, $localAlgorithm$'s execution time increases also with the number of incoherent parts, which depends on the incoherences number, as shown in section IV-C3.

Example 16 (About $localAlgorithm$'s execution time): In this example, we took the object set corresponding to appellation "Leroux, Alain" and the criteria set of example 14. For each tested percentage $randomDate'$:

- we took the 25 $randomDate'$ criteria generated in example 14 and measured each time the execution time of $localAlgorithm$, which is represented by a grey dot on figure 3;
- we calculated the average mark of those 25 execution time measures and represented it by a black dot on on figure 3.

The execution times monotonously increases except for 5 to 15% of random comparison values, from 2 milliseconds on average for 0% to 14 milliseconds on average for 100%.

The spike on 5% to 15% corresponds exactly to the increased number of incoherent parts shown on figure 14. We also notice that execution time fluctuates a lot on these percentage, but fluctuates slightly for other percentages.

For a given number of incoherent parts, increasing the number of incoherences increases the execution time for $localAlgorithm$**.**

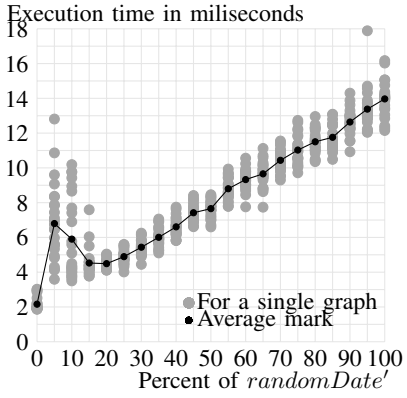

Fig. 3. Execution time to find all best partition values according to local semantics and 8 criteria including $randomDate'$

As a conclusion of this section, **the number of incoherences increases with data ambiguity and decreases with criteria accuracy. The number of incoherences greatly influences algorithms execution time and whether the human partition has a best partition value.**

V. Discussion

Before concluding the paper, let us position our approach with respect to other Entity Resolution methods in the litterature.

A. The entity resolution problem

The entity resolution problem [4][23][20][8][18][21][24] is the problem of identifying as equivalent two objects representing the same real-world entity. This problem is described by many names (entity resolution[4], record linkage[23], reference reconciliation[20], entity matching[8], record matching[18], name disambiguation[21],

data interlinking[24]). The courses of such mismatch can be homonyms (as in people with the same name), errors that occurred at data entry (like "Léa Guizo" for Léa Guizol"), missing attributes (e.g publication date = XXXX), abbreviations ("L. Guizol") or attributes having different values for two objects representing the same entity (change of adress).

The Entity Resolution problems can be adressed as a rule based pairwise comparison rule approach. It is difficult to manually create those rules. Approaches have been proposed in literature [6][10][16] using a training pairs set. Rules can be chained using different constraints: transitivity [3], exclusivity [16] and functional dependencies [1][12][7].

An alternative method for Entity Resolution problem is partitioning (hierarchical partitioning [5], closest neighbour-based method [9] or correlation clustering [3]). Our work falls in this category. The approach closest to our semantics are [3] and [2]. We distinguish ourselves from [3] and [2] because:

- the lack of $neutral$ values in these approaches, and
- the numericalization of symbolic values (numerically aggregated into -1 and $+1$ values), and the use of numerical aggregation methods on these values.

B. Conclusion

In this paper we presented a qualitative investigation on the input and output of symbolic partitioning semantics employed in a real world scenario of an Entity Resolution problem. The lessons learnt from the results presented in this paper are that:

- for a given criteria set, increasing data ambiguity increases the number of incoherences.
- for a given object set, decreasing criteria accuracy increases the number of incoherences.
- increasing the number of incoherences increases the execution time of algorithms used to find all best partitions (except for local semantics when it can also decrease the number of incoherent parts).
- increasing the number of incoherences can make human partition not have a best partition value.

In the aim to detect erroneous links in a bibliographic knowledge base, the presented methodology requires that the human partition has a best partition value. So, global and local semantics must be used according to a set of criteria as accurate as possible to be able to improve chances of the human partition to have a best value.

REFERENCES

[1] R. Ananthakrishna, S. Chaudhuri, and V. Ganti. Eliminating fuzzy duplicates in data warehouses. In *Proceedings of the 28th international conference on Very Large Data Bases*, VLDB '02, pages 586–597. VLDB Endowment, 2002.

[2] A. Arasu, C. Ré, and D. Suciu. Large-scale deduplication with constraints using dedupalog. In *Proceedings of the 25th International Conference on Data Engineering (ICDE)*, pages 952–963, 2009.

[3] N. Bansal, A. Blum, and S. Chawla. Correlation clustering. volume 56, pages 89–113. Springer, 2004.

[4] I. Bhattacharya and L. Getoor. *Entity Resolution in Graphs*, pages 311–344. John Wiley & Sons, Inc., 2006.

[5] M. Bilenko, S. Basil, and M. Sahami. Adaptive product normalization: Using online learning for record linkage in comparison shopping. In *Data Mining, Fifth IEEE International Conference on*, pages 8–pp. IEEE, 2005.

[6] M. Bilenko and R. J. Mooney. Adaptive duplicate detection using learnable string similarity measures. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 39–48. ACM, 2003.

[7] P. Bohannon, W. Fan, F. Geerts, X. Jia, and A. Kementsietsidis. Conditional functional dependencies for data cleaning. In *Data Engineering, 2007. ICDE 2007. IEEE 23rd International Conference on*, pages 746–755. IEEE, 2007.

[8] P. Bouquet, H. Stoermer, and B. Bazzanella. An entity name system (ens) for the semantic web. In *Proceedings of the 5th European semantic web conference on The semantic web: research and applications*, ESWC'08, pages 258–272, Berlin, Heidelberg, 2008. Springer-Verlag.

[9] S. Chaudhuri, V. Ganti, and R. Motwani. Robust identification of fuzzy duplicates. In *Data Engineering, 2005. ICDE 2005. Proceedings. 21st International Conference on*, pages 865–876. IEEE, 2005.

[10] M. Cochinwala, V. Kurien, G. Lalk, and D. Shasha. Efficient data reconciliation. *Information Sciences*, 137(1):1–15, 2001.

[11] M. Croitoru, L. Guizol, and M. Leclère. On Link Validity in Bibliographic Knowledge Bases. In *IPMU'2012: 14th International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems*, volume Advances on Computational Intelligence, pages 380–389, Catania, Italie, July 2012. Springer.

[12] W. Fan. Dependencies revisited for improving data quality. In *Proceedings of the twenty-seventh ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, pages 159–170. ACM, 2008.

[13] A. Guénoche. Partitions optimisées selon différents critères: évaluation et comparaison. *Mathématiques et sciences humaines. Mathematics and social sciences*, (161), 2003.

[14] L. Guizol, M. Croitoru, and M. Leclere. Aggregation semantics for link validity. *AI-2013: Thirty-third SGAI International Conference on Artificial Intelligence*, page to appear, 2013.

[15] L. Guizol, M. Croitoru, and M. Leclère. Aggregation semantics for link validity: technical report. Technical report, LIRMM, INRIA Sophia Antipolis, http://www.lirmm.fr/~guizol/AggregationSemanticsforLinkValidity-RR.pdf, 2013.

[16] R. Gupta and S. Sarawagi. Answering table augmentation queries from unstructured lists on the web. *Proceedings of the VLDB Endowment*, 2(1):289–300, 2009.

[17] M. A. Hernández and S. J. Stolfo. The merge/purge problem for large databases. *SIGMOD Rec.*, 24(2):127–138, May 1995.

[18] M.-Y. Kan and Y. F. Tan. Record matching in digital library metadata. *Commun. ACM*, 51:91–94, February 2008.

[19] A. McCallum, K. Nigam, and L. H. Ungar. Efficient clustering of high-dimensional data sets with application to reference matching. In *Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '00, pages 169–178, New York, NY, USA, 2000. ACM.

[20] F. Saïs, N. Pernelle, and M.-C. Rousset. Reconciliation de references : une approche logique adaptee aux grands volumes de donnees. In *EGC*, pages 623–634, 2007.

[21] N. R. Smalheiser and V. I. Torvik. *Annual Review of Information Science and Technology (ARIST)*, volume 43, chapter Author Name Disambiguation. Information Today, Inc, 2009.

[22] S. Wang. Existence of a pareto equilibrium. *Journal of Optimization Theory and Applications*, 79(2):373–384, 1993.

[23] W. E. Winkler. Overview of record linkage and current research directions. Technical report, BUREAU OF THE CENSUS, 2006.

[24] S. Wölger, C. Hofer, K. Siorpaes, S. Thaler, E. Simperl, and T. Bürger. Interlinking data - approaches and tools. Technical report, STI Innsbruck, University of Innsbruck, 2011.