An Analysis of the SUDOC Bibliographic Knowledge Base from a Link Validity Viewpoint

Léa Guizol, Olivier Rousseaux, Madalina Croitoru, Yann Nicolas, Aline Le Provost

An analysis of the SUDOC bibliographic knowledge base from a link validity viewpoint

Léa Guizol*, Olivier Rousseaux**, Madalina Croitoru*,
Yann Nicolas**, Aline Le Provost**

*LIRMM (University of Montpellier II & CNRS), INRIA Sophia-Antipolis, France
**ABES, France

Abstract. In the aim of evaluating and improving link quality in bibliographical knowledge bases, we develop a decision support system based on partitioning semantics. The novelty of our approach consists in using symbolic values criteria for partitioning and suitable partitioning semantics. In this paper we evaluate and compare the above mentioned semantics on a real qualitative sample. This sample is issued from the catalogue of French university libraries (SUDOC), a bibliographical knowledge base maintained by the University Bibliographic Agency (ABES).

1 Introduction

Real World Context. The SUDOC (catalogue du Système Universitaire de Documentation) is a large bibliographical knowledge base managed by ABES (Agence Bibliographique de l'Enseignement Supérieur). The SUDOC contains **bibliographic notices** (document descriptions $\approx 10.000.000$), and **authorship notices** (person descriptions $\approx 2.000.000$). An authorship notice possesses some attributes (ppn[1], appellation set, date of birth...). A bibliographic notice also possesses some attributes (title, ppn[1], language, publication date...) and **link**(s) to authorship notices. A link is labeled by a **role** (as *author*, *illustrator* or *thesis advisor*) and means that the person described by the authorship notice has participated as the labeled role to the document described by the bibliographic notice.

One of the most important tasks for ABES experts is to reference a new book in SUDOC. To this end, the expert has to register the title, number of pages, types of publication domains, language, publication date, and so on, in a new bibliographic notice. This new bibliographic notice represents the physical books in the librarian hands which he/she is registering. He/she also has to register people which participated to the book's creation (namely the **contributors**). In order to do that, for each contributor, he/she selects every authorship notice (named *candidates*) which has an appellation similar to the book contributor. Unfortunately, there is not that much information in authorship notices because the librarian politics is to give minimal information, solely in order to distinguish two authorship notices which have the same appellation, and nothing more (they reference books, not people!). So the librarian has to look at bibliographic notices which are linked to authorship notices candidates (the *bibliography* of candidates) in

[1] A ppn identifies a notice.

order to see whether the book in his/her hands seems to be a part of the bibliography of a particular candidate. If it is the case, he/she links the new bibliographic notice to this candidate and looks at the next unlinked contributor. If there is no good candidate, he/she creates a new authorship notice to represent the contributor.

This task is fastidious because it is possible to have a lot of candidates for a single contributor (as much as 27 for a contributor named "BERNARD, Alain"). This creates errors, which in turn can create new errors since linking is an incremental process. In order to help experts to repair erroneous links, we proposed two **partitioning semantics** in [11] which enables us to detect erroneous links in bibliographic knowledge bases. A partitioning semantics evaluates and compares **partitions**[2].

Contribution. The contribution of this paper is to practically evaluate the results quality of partitioning semantics [11] on a real SUDOC sample. We recall the semantics in section 3, clearly explain on which objects and with which criteria the semantics have been applied in section 2, and present qualitative results in section 4. We discuss the results and conclude the paper in section 5.

2 Qualitative experiments

In this section, we first adapt the **entity resolution problem**[3][4] to investigate link quality in SUDOC in section 2.1. This problem is known in literature under very different names (as record linkage [18], data deduplication [2], reference reconciliation [16]...). Then we define (section 2.3) and detail (section 2.4) criteria used in order to detect erroneous links in SUDOC. Those criteria are used on SUDOC subsets defined in section 2.2.

2.1 Contextual entities: from erroneous links to entity resolution

In order to detect and repair erroneous links, we represent SUDOC links into **contextual entity** (the i contextual entity is denoted Nc_i). A contextual entity represents a bibliographic notice Nb_j from the viewpoint of one of its contributor, named the C contributor of Nc_i and denoted $C(Nc_i)$. The contextual entities are compared together with an entity resolution method, in order to see which ones have a contributor representing a same real-word person. As explained in [8], traditional entity resolution methods cannot be directly applied. This entity resolution method is supposed to group (in a same class of the created partition) the contextual entities such as their C contributor represents a same real-word person, and to separate the other ones (to put them in distinct partition classes). A contextual entity Nc_i has several attributes. Most of them are $Nb(Nc_i)$ attributes (as title, publication date, publication language, publication domain codes list) and others depend on the C contributor:

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

[3] The **entity resolution problem** is the problem of identifying as equivalent two objects representing the same real-world entity.

– role of the C contributor (there is a set of typed roles as “thesis_advisor”),
– list of the possible appellations of the C contributor. An appellation is composed of a name and a surname, sometimes abbreviated (as “J.” for surname),
– list of contributors which are not C. For each of them, we have the identifier of the authorship notice which represents it, and the role.

The publication language attribute is typed (for example, “eng” for English language, “fre” for French language and so on). The publication date is most of the time the publication year (“1984”). Sometimes information is missing and it only gives the century or decade (“19XX” means that the document has been published last century). A publication domain is not a describing string but a code with 3 digits which represent a domain area.

Example 1 (Contextual entity attributes). The authorship notice of ppn **026788861**, which represents “CHRISTIE, Agatha” is linked as “author” to the bibliographic notice of ppn 121495094, which represents “*Evil under the sun*” book. The contextual entity which represents this links has the following attributes:

– title: “*Evil under the sun*”
– publication date: “2001”
– publication language: “eng”
– publication domain codes list: $\{\}$ (they have not been given by a librarian)
– list of the possible appellations of the C contributor: {“CHRISTIE, Agatha”,“WESTMACOTT, Mary”,“MALLOWAN, Agatha”,“MILLER, Agathe Marie Clarissa”}
– role of the C contributor: “author”
– list of contributors which are not C: $\{\}$ (there is no other contributors in this case)

Let Nc_i be the contextual entity identified by i. As any contextual entity, it has been constructed because of a link between an authorship notice and a bibliographic notice, which are respectively denoted $Na(Nc_i)$ and $Nb(Nc_i)$. We define two particular partitions: the initial one and the human one.

The **initial partition** (denoted Pi) of contextual entities is the one such as two contextual entities Nc_i, Nc_j are in a same class if and only if $Na(Nc_i) = Na(Nc_j)$. This represents the original organization of links in SUDOC.

The **human partition** (denoted Ph) of contextual entities is a partition based on an human expert's advice: two contextual entities Nc_i, Nc_j are in a same class if and only if the expert thinks that their C contributor corresponds to a same real word person.

The goal of this paper's work is to distinguish SUDOC subsets constructed as in the following section 2.2 with or without erroneous links. We make the hypothesis that the human partition has to be a best one (because it is the good one according to expert) and that the initial partition has to not be a best partition except if $Pi = Ph$. So, partitioning semantics are approved if Ph is a best partition according to the semantics, but not Pi. Let us determine what is a SUDOC contextual entities subset to partition.

2.2 Selecting contextual entities on appellation

A SUDOC subset $\mathbb{O}$ selected for an appellation A contains all contextual entities which represent a link between any SUDOC bibliographic notice and a SUDOC authorship

notice which has an appellation close to the appellation A. To select a SUDOC subset for a given appellation (as "BERNARD, Alain") is a way to separate SUDOC in subsets which can be treated separately, as the canopies [15] and blocking[13] methods does. This is also a simulation of how experts select a SUDOC subset to work on it, as explained in part 1. In the following, we will only be interested into partitioning SUDOC subsets selected for an appellation. Let us define and describe criteria used in order to compare contextual entities together.

2.3 Symbolic criteria

In the general case, a **criterion** is a function which compares two objects and returns a comparison value. Let c be a criterion, and o_i, o_j are two objects. We denote $c(o_i, o_j)$, the comparison values according to c between o_i and o_j.

In this work case, we use **symbolic criteria** which can return $always$, $never$, $neutral$, a **closeness value** or a **farness value** as comparison value. $always$ (respectively $never$) means that objects have to be in a same (respectively distinct) partition class[2]. Closeness (respectively farness) values are more or less intense and far from the $neutral$ value, meaning that objects should be in a same (respectively distinct) partition class. Closeness (respectively farness) values are strictly ordered between themselves, specific to a criterion and less intense than $always$ (respectively $never$). Those values are denoted $+$,$++$ and so on (respectively $-$,$--$) such as the more $+$ (respectively $-$) symbols they have, the more intense and the further from $neutral$ the value is. For a criterion, $always$ is more intense than $+++++$, which is more intense than $++$ which is more intense than $+$. $+$ is only more intense than $neutral$. $neutral$ means that the criterion has no advice about whether to put objects in a same class or not.

2.4 Criteria for detecting link issues in SUDOC

In order to simulate human expert behaviour, nine symbolic criteria have been developed. Some are closeness-criteria[4] ($title$, $otherContributors$), farness-criteria[4] ($thesis$, $thesisAdvisor$, $date$, $appellation$, $language$) and others are both ($role$, $domain$). Each of these criteria give the $neutral$ comparison value when a required attribute of a compared contextual entity is unknown and by default. Let Nc_i, Nc_j be two contextual entities.

- $appellation$ criterion is a particular farness-criterion. Indeed, it compares appellation lists to determine which contextual entities can not have a same contributor C. When it is certain (as when appellations are "CONAN DOYLE, Arthur" and "CHRISTIE, Agatha"), it gives a $never$ comparison value, which forbids other criteria to compare the concerned authorship notices together. This is also used to divide SUDOC in subsets which should be evaluated separately.
- $title$ criterion is a closeness-criterion. This criterion can give an $always$ value and 3 closeness comparison values. It is based on a Levenshtein comparison [14]. It is

[4] A closeness-criterion (respectively a farness-criterion) c is a criterion which can give a closeness or $always$ (respectively a farness or $never$) comparison value to two objects.

useful to determine which contextual entities represent a same work, edited several times. This is used by the $thesis$ criterion.

- $otherContributors$ criterion is a closeness-criterion. It counts the others contributors in common, by comparing their authorship notices. One (respectively several) other common contributor gives a $+$ (respectively $++$) comparison value.
- $thesis$ criterion is a farness-criterion. $thesis(Nc_i, Nc_j) = -$ means that Nc_i, Nc_j are contextual entities which represent distinct thesis (recognized thanks to the $title$ criterion) from their "author" point of view. $thesis(Nc_i, Nc_j) = --$ means that Nc_i, Nc_j have also been submitted simultaneously.
- $thesisAdvisor$ criterion is a farness-criterion. $thesisAdvisor(Nc_i, Nc_j) = --$ (respectively $-$) means that Nc_i and Nc_j have a same contributor C if and only if this contributor has supervised a thesis before (respectively two years after) submitting his/her own thesis.
- $date$ criterion is a farness-criterion. For 100 (respectively 60) years at least between publication dates, it gives a $--$ (respectively $-$) comparison value.
- $language$ criterion is a farness-criterion. When publication languages are distinct and none of them is English, $language$ returns a $-$ value.
- $role$ criterion returns $+$ when contributor C roles are the same (except for current roles as "author", "publishing editor" or "collaborator"), or $-$ when they are distinct (except for some pairs of roles as "thesis advisor" and "author").
- $domain$ criterion compares list of domain codes. Domain codes are pair-wise compared. $domain(Nc_i, Nc_j)$ gives closeness (respectively farness) comparison values if every Nc_i domain codes is close (respectively far) from a Nc_j domain code and the other way around.

Let us resume global and local semantics before to evaluate their relevance with respect to the above mentioned criteria on real SUDOC subsets.

3 Partitioning semantics

Let us summarize partitioning semantics detailed in [11]. A partitioning semantics evaluates and compares partitions on a same object set. The following partitioning semantics (in sections 3.1 and 3.2) are based on symbolic criteria.

3.1 Global semantics

In this section we define what is a a best partition on the object set $\mathbb{O}$ (with respect to the $\mathbb{C}$ criteria set) according to global semantics. A partition has to be **valid**[5][2] in order to be a best one. A partition P has also an **intra value** and an **inter value** per criterion of $\mathbb{C}$. The intra value of a criterion c depends of the most intense (explained in section 2.3) farness or $never$ value of c such as it compares two objects in a same

[5] A partition P is **valid** if and only if there is no two objects o_i,o_j such as: (i) they are in a same class of P and they $never$ have to be together according to a criterion (expressed by $never$ comparison value), or (ii) they are in distinct P classes but $always$ have to be together according to at least a criterion.

class (should not be the case according to c). In the same way, the inter value of c depends of the most intense closeness or $always$ 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 [10]. We note that the $neutral$ comparison value does not influence partition values.

A partition P on an object set $\mathbb{O}$ is a best partition according to a criteria set $\mathbb{C}$ if P is valid and P has a best value, meaning that it is impossible to improve an inter or intra value of any criterion $C \in \mathbb{C}$ without decreasing inter or intra value of a criterion $C' \in \mathbb{C}$ (it is a Pareto equilibrium [17]).

id	title	date	domains [...]	appellations
Nc_1	"*Letter to a Christian nation*"		religion	"HARRIS, Sam"
Nc_2	"*Surat terbuka untuk bangsa kristen*"	2008	religion	"HARRIS, Sam"
Nc_3	"*The philosophical basis of theism*"	1883	religion	"HARRIS, Samuel"
Nc_4	"*Building pathology*"	2001	building	"HARRIS, Samuel Y."
Nc_5	"*Building pathology*"	1936	building	"HARRIS, Samuel Y."
Nc_6	"*Aluminium alloys 2002*"	2002	physics	"HARRIS, Sam J."

Table 1. Example of objects set

Example 2 (Global semantics evaluating a partition on an object set $\mathbb{O}$).

Let us represent an object set $\mathbb{O} = \{Nc_1, Nc_2, Nc_3, Nc_4, Nc_5, Nc_6\}$ in table 1. Each object is a contextual entity and represents a link between a bibliographic notice and an authorship notice (here, an "author" of a book). Id is the object identity. For each of them, title, date of publication, publication domain and appellation of the contributor C are given as attributes.

Nc_1 and Nc_2 represent a same person, as Nc_4, Nc_5 does. The human partition on $\mathbb{O}$ is: $Ph = \{\{Nc_1, Nc_2\}, \{Nc_3\}, \{Nc_4, Nc_5\}, \{Nc_6\}\}$. This partition, according to global semantics and with respect to the criteria set $\mathbb{C} = \{appellation, title, domain, date\}$ (criteria are detailed in section 2.4) is not coherent with some of $\mathbb{C}$ criteria. The Ph 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 publication dates distant of more than 60 years and less than 100 years.

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

3.2 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 parts of $\mathbb{O}$. Parts of $\mathbb{O}$ can be coherent or incoherent. An **incoherent 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 $Nc_i \in \mathbb{O} - \mathbb{O}_a$, $Nc_j \in \mathbb{O}_a$, and $c \in \mathbb{C}$;
- there is no subset of $\mathbb{O}_a$ for which the previous property is true;
- there is $b(o_k, o_l)$, a farness or $never$ value such as $o_k, o_l \in \mathbb{O}_a$, and $b \in \mathbb{C}$.

An *incoherent part partition value* is based on every comparison between objects which are in it. The **coherent part** of an object set $\mathbb{O}$ is a $\mathbb{O}$ subset containing every $\mathbb{O}$ object which is not in a incoherent part of $\mathbb{O}$. The *coherent part partition value* of $\mathbb{O}$ is based on every comparison between objects which are not in the same incoherent part of $\mathbb{O}$.

Example 3 (Incoherent and coherent parts).

Let us identify incoherent parts of the object set $\mathbb{O}$ according to $\mathbb{C}$ given in example 2. Nc_1, Nc_2, Nc_3 are close together due to $domain$ criterion: they are about religion. Nc_1, Nc_2, Nc_3 are not close to Nc_4, Nc_5 or Nc_6 according to any of $\mathbb{C}$ criteria and Nc_2, Nc_3 are far according to $date$ criterion ($date(Nc_2, Nc_3) = --$) so $\{Nc_1, Nc_2, Nc_3\}$ is an incoherent part of $\mathbb{O}$. 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 are also far according to $date$ criterion ($date(Nc_4, Nc_5) = -$) so $\{Nc_4, Nc_5\}$ is also an incoherent part.

So, there are 2 incoherent parts in $\mathbb{O}$: $\{Nc_1, Nc_2, Nc_3\}$ and $\{Nc_4, Nc_5\}$. Nc_6 is not in an incoherent part so Nc_6 is in the coherent part of $\mathbb{O}$.

A partition on $\mathbb{O}$ is a *best partition according to local semantics* if it has best partition values for each incoherent part of $\mathbb{O}$ and for the $\mathbb{O}$ coherent part.

Example 4 (Local semantics evaluating a partition on an object sets $\mathbb{O}$).

In example 3, we identified the incoherent parts of the object set $\mathbb{O} = \{Nc_1, Nc_2, Nc_3, Nc_4, Nc_5, Nc_6\}$ according to the criteria set $\mathbb{C} = \{appellation, title, domain, date\}$.

The partition on $\mathbb{O}$ given in example 2: is $Ph = \{\{Nc_1, Nc_2\}, \{Nc_3\}, \{Nc_4, Nc_5\}, \{Nc_6\}\}$. According to local semantics, Ph has 3 values, one for the coherent part and 2 for incoherent parts (1 per incoherent part):

- a perfect value for the coherent part of $\mathbb{O}$;
- the incoherent part $\{Nc_1, Nc_2, Nc_3\}$ has a very bad inter value for the $domain$ criterion ($always$);
- the incoherent part $\{Nc_4, Nc_5\}$ has an bad intra value for the $date$ criterion ($--$);

This semantics enables us to split an object set into several parts which can be evaluated separately. We explained local and global semantics in this part, which are a way to solve the entity resolution problem. Let us evaluate them on a real SUDOC sample.

4 Results

ABES experts have selected 537 contextual entity divided into 7 SUDOC subsets selected for an appellation. The table 2 shows for each SUDOC subset selected for an appellation A (please see section 2.2):

1. $|Nc|$ is the number of contextual entities which represent a link between a bibliographic notice and an authorship notice which has a close appellation to A,
2. $|Na|$ is the number of authorities notices according to human partitions (corresponding to class number of human partition),
3. "Ph best" (respectively "Pi best") shows whether the human partition Ph (respectively initial partition Pi) has a best value according to global semantics and with respect to all 9 criteria detailed in part 2.4,
4. $Ph \succ Pi$ is true if and only if Ph has a better value than Pi.

Appellation	$\|Nc\|$	$\|Na\|$	Ph best	Pi best	$Ph \succ Pi$	Ph' best	Repairs
"BERNARD, Alain"	165	27	no	not valid	yes	yes	
"DUBOIS, Olivier"	27	8	no	no	yes	no	1
"LEROUX, Alain"	59	6	no	not valid	yes	yes	
"ROY, Michel"	52	9	yes	not valid	yes	yes	
"NICOLAS, Maurice"	20	3	yes	no	yes	yes	
"SIMON, Alain"	63	13	no	no	yes	no	1
"SIMON, Daniel"	151	16	no	not valid	yes	yes	

Table 2. Human and initial partitions with respect to 9 criteria and global semantics

Local semantics, has the same results than global semantics on this sample.

For global semantics, Pi is never a best partition. 5 times out of 7, Ph does not have a best value (each time, it is due to the $domain$ and $language$ criteria, and two times $thesisAdvisor$ is also involved), but it is all the time valid and better than Pi, which is encouraging for erroneous link detection. Erroneous links are particularly obvious when Pi is not even valid (4 times out of 7). It is due to the $title$ criterion detailed in part 2.4. We regret that Ph is not all the time a best partition, but the global semantics is able to distinguish Pi from Ph in 5 cases out of 7: when Pi is not valid, or when Ph is a best partition but not Pi.

Because the $domain$ and $language$ criteria often considers that Ph is not a good enough partition, Ph was also evaluated for global semantics according to all criteria without $domain$ and $language$ (shown in table 2 in column "Ph' best") and that increases the human partition which obtains a best value in 3 more cases (for "BERNARD, Alain", "SIMON, Daniel" and "LEROUX, Alain" appellations). This tells us that $domain$ and $language$ criteria are not reasonably accurate.

In order to evaluate if Ph is far from having a best partition value, we enumerate the number of repairs to transform Ph' into a partition Ph'' which has a best value according to all criteria except $domain$ and $language$. We show this repair number in the "Repairs" column of table 2. An atomic repair could be:

– merging two partition classes (corresponds to merging two contextual entities which represent a same real word person), or

– splitting a partition class in two classes (corresponds to separate books which are attributed to a same real word person but belong to two distinct real word persons).

We can see that only a few repairs are needed compared to the number of classes (corresponding to $|Na|$ column in the table): 1 repair for "DUBOIS, Olivier" and for "BERNARD, Alain" appellations.

Let us highlight that observing human partition values has permitted to *detect and correct an erroneous link* (for "ROY, Michel" appellation) in the human reference set, validated with experts. The global semantics does not always consider that the human partition is a best partition, but in the worst case the human partition is very close to be one according to repairs number, and global semantics allow us to detect that initial partitions are much worse than human partitions. This last point is encouraging. This means that the semantics can also be useful to help in criteria tuning, by showing which criteria are bad according to human partitions, and for which authorship notices comparison. For example, the fact that the human partition value is often bad according to the $domain$ criterion shows that this criterion is actually not an accurate criterion. Let us talk about other entity resolution methods and conclude.

5 Discussion

The entity resolution problem [4][18][16][6] is the problem of identifying as equivalent two objects representing the same real-world entity. The causes of such mismatch can be due to 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 address).

The entity resolution problem can be addressed as a rule based pairwise comparison rule approach. Approaches have been proposed in literature [12] using a training pairs set for learning such rules. Rules can be then be chained using different constraints: transitivity [3], exclusivity [12] and functional dependencies [1] [9].

An alternative method for entity resolution problem is partitioning (hierarchical partitioning [5], closest neighbor-based method [7] or correlation clustering [3]). Our work falls in this last category. Due to the nature of treating criteria values, the closest approach to our semantics are [3] and [2]. We distinguish ourself to [3] and [2] because of the lack of $neutral$ values in these approaches, the numerization of symbolic values (numerically aggregated into -1 and $+1$ values), and the use of numerical aggregation methods on these values.

Conclusion. In this paper we proposed a practical evaluation of the global and local semantics proposed in [11]. The conclusions of this evaluation are:

– For SUDOC subsets selected by appellation, *both semantics* are effective to distinguish a human partition from the initial partition; however it is not perfect with respect to our set of criteria (if the human partition is not a best partition, it has a close value).
– *Both semantics* could be useful to detect meaningless criteria.

As immediate next steps to complete this our work we mention using global or local semantics to improve implemented criteria.

Acknowledgements 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.

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. 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.
7. 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.
8. 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.
9. 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.
10. A. Guénoche. Partitions optimisées selon différents critères: évaluation et comparaison. *Mathématiques et sciences humaines. Mathematics and social sciences*, (161), 2003.
11. 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.
12. 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.
13. M. A. Hernández and S. J. Stolfo. The merge/purge problem for large databases. *SIGMOD Rec.*, 24(2):127–138, May 1995.
14. V. I. Levenshtein. Binary codes capable of correcting deletions, insertions and reversals. In *Soviet physics doklady*, volume 10, page 707, 1966.
15. 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.

16. 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.
17. S. Wang. Existence of a pareto equilibrium. *Journal of Optimization Theory and Applications*, 79(2):373–384, 1993.
18. W. E. Winkler. Overview of record linkage and current research directions. Technical report, BUREAU OF THE CENSUS, 2006.