Identification of FRBR Works Within Bibliographic Databases: An Experiment with UNIMARC and Duplicate Detection Techniques

Nuno Freire, José Borbinha, and Pável Calado

INESC-ID, Rua Alves Redol 9, Apartado 13069
1000-029 Lisboa, Portugal
nuno.freire@ist.utl.pt, jlb@ist.utl.pt,
pavel.calado@tagus.ist.utl.pt

Abstract. Many experiments and studies have been conducted on the application of FRBR as an implementation model for bibliographic databases, in order to improve the services of resource discovery and transmit better perception of the information spaces represented in catalogues. One of these applications is the attempt to identify the FRBR work instances shared by several bibliographic records. In our work we evaluate the applicability to this problem of techniques based on string similarity, used in duplicate detection procedures mainly by the database research community. We describe the particularities of the application of these techniques to bibliographic data, and empirically compare the results obtained with these techniques to those obtained by current techniques, which are based on exact matching. Experiments performed on the Portuguese national union catalogue show a significant improvement over currently used approaches.

Keywords: Functional Requirements for Bibliographic Records, FRBR, Bibliographic databases, string similarity, duplicate detection.

1 Introduction

FRBR - Functional Requirements for Bibliographic Records [1], is a conceptual model developed by the IFLA - International Federation of Library Associations and Institutions, proposing how the bibliographic information should be represented. A purpose of the FRBR model is to bring the bibliographic world closer to the actual multimedia, digital and more heterogeneous world. Information systems supporting information schemas compatible with the FRBR model are supposed to assure e richer representation of the information, and therefore, to provide better services of resource discovery and transmit better perception of the information spaces represented in catalogues.

FRBR is widely recognized as a valuable model, but with an important constraining: as it represents a richer semantic model, it is not easy to "upgrade" to its level the existing bibliographic catalogues (it is not trivial to extract new semantics concepts from information structures that were not designed to hold them…). Considering the millions of MARC related records existing nowadays all over the works, created at a

D.H.-L. Goh et al. (Eds.): ICADL 2007, LNCS 4822, pp. 267–276, 2007.

very high cost, made the problem of converting traditional catalogues to FRBR structures a very relevant one (while on the same time very challenging).

This is the main motivation for the work reported in this paper, where we try to contribute to the solution of the problem of building FRBR structures from inherited UNIMARC bibliographic records by applying techniques of detection of duplicate information.

The detection of duplicate records is an area of great interest in the traditional database research community. It is a problem faced many times when implementing a data warehouses or any other system that aggregates data from heterogeneous data sources. Often, the same entities in the real world have two or more representations in such databases. These duplicate records do not share a common key, they may differ in structure and in lexicon, and they may even contain errors, making their detection a very difficult task.

In fact, this problem is common to many research communities, although the term used is not always the same [1]: record linkage, record matching, merge-purge, data deduplication, instance identification, database hardening, and name matching. In our work we evaluated the applicability of some of the techniques developed in these communities to the purpose of detecting common expressions of works within bibliographic databases, according to the FRBR definition.

This paper follows with a description of the main techniques for the detection of duplicate records. After that we analyse the problem of the detection of FRBR works in groups of UNIMARC records, and we formulate our hypothesis to address it. In the following section we describe the experiment designed for this purpose, as also the results achieved. The paper continues with a discussion of the results, a description of related work, and finished with the conclusions and references to future work.

2 Techniques for the Detection of Duplicate Records

When setting up a duplicate detection process, several issues have to be addressed. Which data will be used for comparison, how should it be coded, how fields are to be matched individually, and on what conditions the comparison results of the individual fields identify two records as duplicates. In the rest of this section we describe these issues and the general approaches widely used to solve them.

The process starts with the preparation of data for further processing. This step comprises tasks for selecting the relevant data from the data sources, and parsing and transforming it so that in conforms to a standardized data schema. Data preparation greatly reduces the structural heterogeneity of the source data, but misspellings and different conventions for recording the same information continue to result in different, multiple representations of a unique object in the database. For this reason, records have to be compared for their similarity, by measuring similarities of the fields, one by one. These similarity results can then be further processed to decide if the records match.

One of the main obstacles to the detection of duplicates is the typographical variations of string data. Therefore string comparison techniques have bean a very active topic in research. Among the many techniques for matching string fields, three main types of similarity metrics can be considered: character based, token based and

phonetic. How well each of these techniques works on evaluating the similarity between strings depends on the characteristics of the data they are applied to.

Character based similarity metrics are most suitable to handle typographical errors. They measure the amount of edit operations (insert, delete, replace) that are necessary to transform one string into the other being compared. These techniques don't work well in cases where typographical conventions lead to rearrangement of words (people's names, for example, may be entered by their surname or by their first name). Token-based metrics, on the other hand, try to compensate for this problem by matching tokens in the strings (typically words) independently of their location within the strings. These techniques usually make use of token-weighting schemes, such as the "term frequency–inverse document frequency" (TFIDF) [2]. Finally, phonetic techniques address those cases where strings may be phonetically similar even if they are not similar at character or token level. These are widely used to match fields containing person surnames.

Several algorithms exist for all these kinds of metrics and the decision of which field comparison techniques to use is not an easy one. Analysis of the few existing studies seams to indicate that no single metric is suitable for all data sets [1, 4]. In many cases, using flexible metrics that can accommodate multiple similarity comparisons may lead to the best results.

The final decision to match two records is made by reasoning on the similarity scores obtained by comparing the individual fields. Depending on the complexity of the record structure and on the possibility of creating a training set of data, two types of techniques may be implemented: declarative techniques or machine learning techniques. In general, better results can be obtained by using machine learning techniques [1, 6].

3 Detection of FRBR Works in UNIMARC Records

Previous experiments on the identification of works within bibliographic databases have not fully explored the applicability of duplicate detection techniques. The methods deployed usually consist of algorithms that have a strong emphasis on the data preparation phase, to create keys that identify the work from data in the bibliographic record. These keys are then used to match the records using declarative techniques, with decision tables or sets of rules [7, 8]. However, the fields that make up the keys are compared using exact comparison without, resorting to similarity metrics. Although some of the heterogeneity of data is handled quite well in the data preparation phase, a simple exact comparison may be insufficient. This causes record matches to be missed.

The two most important fields for matching works in UNIMARC records are the title (including subtitles) and the authors. Both fields are prone to typing errors, and the use of abbreviations is frequent, which is enough to disable exact matching of records (two examples are shown in Table 1). Additionally, in PORBASE, we observed that subtitles may be recorded in different forms: some are omitted, others are recorded separately from the title, others together with the title, and, sometimes, in records with more than one subtitle, they may be recorded in different orders.

The authors' field may not be as problematic as titles, since authority control practices used in libraries share the same cataloguing rules and procedures. However, when

trying to identify works across different data sets from unrelated libraries (especially if these are from different countries) exact matches will be much harder to find for authors, mainly because of different spelling of names across different languages.

From the previous analysis, we formulate the hypotheses that the use of similarity metrics applied to titles and authors could improve the identification of works across existing bibliographic records, by increasing the number of relevant record matches without a significant increase in the number of false matches. To test our hypotheses we designed a set of experiments to compare the use of exact matching techniques to the use of matching based on similarity metrics.

Table 1. Example of two works described in bibliographic records extracted from PORBASE. We can observe the occurrence of typing errors in both title and author fields and different structures for the subtitle field.

Rec.	Title	Subtitles	Authors
A	Grammar's great!	exercícios com soluções, 5º, 6º, 7º anos	Sottomayor, Maria Manuela
B	Grammar's great!	5º, 6º e 7º anos	Sotomayor, Maria Manuela
		exercícios com soluções	
C	Anti-gadouel	français, niveau 6-8, 12ème annéee	Gueidão, Ana
			Crespo, Idalina
D	Anti gadoue	francês	Gueidão, Ana
		niveau 6-8	Crespo, Idalina
		12ème année	

4 Related Work

Few works have explored the computer aided identification of FRBR work entities in bibliographic databases. The major reference works in this area are the several experiments that have been conducted by OCLC [7] and that have been applied by several other projects.

Of particular interest is the approach from the Melvyl Recommender Project [8] that tries to match records, also when titles and authors don't match exactly, by using other data in the bibliographic records, such as dates, identifiers, and by taking in consideration partial matches in author names and subtitles. However we don't know of any other experiment or study that tried to use similarity metrics in this specific task.

Our work has some overlap with the work carried out in citation indexing systems, that autonomously index the citations found in research papers [12, 13, 14]. These works also use similarity metrics to match author names and titles. However, similarity metrics are very data sensitive, and the FRBR work instance identification has patterns of data heterogeneity different from citation matching.

5 The Experiment

The experiments were carried out in two data sets of UNIMARC bibliographic records: (1) PORBASE, the full Portuguese National Bibliographic Database; and (2)

the record set for Porto Editora, a major Portuguese book publisher, which is a subset of records also taken from PORBASE.

The data set from Porto Editora consists of 6.492 records. This publisher is focused on educational works (school manuals, classic works, dictionaries, etc.) which typically have multiple editions, making this data set very appropriate to validate similarity techniques, as we can measure precision and recall. To measure them, we used a similarity metric with a low similarity threshold and manually classified the record matches. This classification was based on a summarized version of the bibliographic records containing only the titles, authors, ISBNs, editions and publication dates.

The PORBASE data set, containing 1.360.686 records, was our real target. It is the largest bibliographic database in Portugal, with collections from nearly 200 different libraries. Our assumption was that once we had our techniques tuned with the Porto Editora set, we could accept the results with a higher level of confidence in PORBASE.

The data preparation phase followed a similar process to the one defined in the OCLC FRBR work-set algorithm [7], now adapted to the UNIMARC format.

5.1 Similarity of the Titles

Some data heterogeneity was still evident after data preparation. Problems such as misspelling and typing errors, lack of spaces between words, abbreviations, various ways of recording subtitles, and missing words would still occur.

A survey on the comparison studies that where conducted on data with similar characteristics indicated that a token based metric should be used [9]. Preliminary experiments were performed to determine the most appropriate similarity metric for titles. A combination of the Jaro-Winkler metric [10] with a TFIDF weighting scheme[1] gave the best results on our preliminary tests. In fact, the results obtained with any single metric were not very satisfactory, which lead to the adoption of combined metrics. Although several metrics resulted in high recall in the matching records, their precision was lower than expected. This came from the fact that small variations in the title of different works from the same authors are very common.

We created a similarity metric that adjusts the similarity score given by Jaro-Winkler-TFIDF to better distinguish between similar titles that refer to the same work from those that do not. We will refer to this metric as BRT metric for the remainder of this paper.

When processing titles, the main problems found were the following. Similarity is significantly lowered when a difference in numeration existed between titles (number, year, roman numerals, and single letter like ‘A’ ‘B’ ‘C’). Examples of these cases where found in school manuals (as, for example, a mathematics manual of different levels by the same authors: “Mathematics 7” and “Mathematics 8”).

Title length greatly influences the similarity values. If two long titles differ in just one or two words, the similarity score will still be high. Therefore, these non-matching words should be compared one by one and, if a word was significantly

[1] An implementation of Jaro-Winkler with TFIDF from the SecondString project (http://secondstring.sourceforge.net/) was used for this purpose.

different from the other (as given by the Smith-Waterman metric), the similarity score was lowered.

Finally, we observed that sometimes some words were omitted in the titles. However, these were typically stop words, i.e., very frequent words that carry very little information, thus the penalty to the similarity score was only marginal.

5.2 Similarity of Author Names

The similarity between author names was measured using the Jaro-Winkler metric. In fact, the only discrepancies found where caused by typing errors or missing or abbreviated middle names, which are easily matched by Jaro-Winkler. For this reason, no further algorithms where tested.

5.3 Final Matching

The final decision to match the records was done by a declarative rule that used, as input, the similarity scores for title and author names. The similarity metrics gave results between 0 (no similarity exists) and 1 (identical). The matching rule defined a minimum threshold of 0,7 for authors, of 0,65 for titles, and of 0,6 for the product of both the similarity scores. The choice of these similarity thresholds was based on the results obtained in our experiments at different thresholds (shown in section 4.5).

Because comparing the similarity of fields is a time consuming process, it is imperative to avoid comparing every record to every other record in the database. To solve this problem, we took a clustering approach. Clusters of the titles were created based on cosine similarity of the titles, and only the records within the same cluster were compared for similarity. This technique improved the performance because the creation of the clusters is much faster than measuring the similarity for all records, reducing the number of record similarity comparisons to a great extent.

5.4 Exact Matching Process

A second independent process to detect duplicate works was implemented without resorting to similarity metrics. The purpose was to compare the results from exact matching to those of similarity matching.

The titles and authors were stored in a relational database after the data preparation phase, which was the same as for both processes. Matching of titles and authors was done using SQL queries and fields would only match if they had exactly the same data.

5.5 The Results

When using the similarity metrics, we obtained the following results. The Jaro-Winkler method identified all cases of similar author names in the test data sample. Matching of authors by similarity accounted for less than 1% of the total matched authors. Further analysis or comparison with other similarity metrics was not performed on authors, since the data set size was too small to draw any meaningful conclusions.

For the titles the sample proved to be an excellent test case, with a very high number of cases that exact comparison missed and were matched by the similarity metrics.

The measured recall for the exact matching process was 63.84%. Figure 1 shows the recall and precision results of three experiments with similarity metrics on the Porto Editora data set. In all three experiments, authors were always compared using the Jaro-Winkler metric, while the similarity metric used for titles varied.

The first test was performed by comparing the titles using the Jaro-Winkler with TFIDF. This metric greatly improved recall, when compared to exact matching, yielding recall values above 98%. Precision was lower than exact matching, as was expected. To obtain gains in recall above 98%, precision started to deteriorate to unsatisfactory levels.

The analysis of the precision failures of the previous method led to the development of the BRT metric that is adapted to the comparison of titles in bibliographic records (as described in section 4.1). This similarity metric obtained the best results of our tests, with both precision and recall levels very close to 100%.

A third test was performed in an attempt to further improve the recall of the previous test. During the experiments, we observed that most of the missed matches where caused by a missing subtitle on one of the records. We therefore adapted the BRT metric by slightly increasing the similarity score in these cases. This change yielded little gain in recall at some similarity thresholds (data not shown) but the best result was still obtained by the BRT metric.

Figure 1 shows the results of the two metrics tested, at 5 levels of similarity thresholds for each metric. Table 2 shows the best results obtained for each method and the similarity threshold at which they were obtained.

We also tested the clustering method used to reduce the number of record similarity calculations necessary. We checked for any missed matches and to what extent it reduced the number of comparisons on the Porto Editora data set. We observed that the number of records comparisons was reduced from 21.063.295 to 782.853 and no records matches where missed.

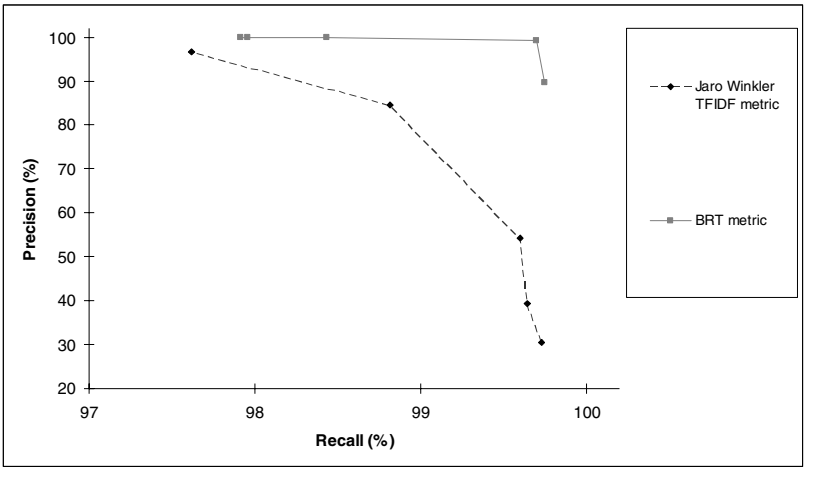

Fig. 1. Recall and precision results of the two methods used on titles, at the most relevant similarity thresholds, in the Porto Editora data set

Table 2. Best recall/precision relation obtained with the three matching methods used for matching titles on the Porto Editora data set

Matching method	Recall (%)	Precision (%)	Similarity threshold
Exact matching	63,84	100,00	-
Jaro Winkler TFIDF metric	98,82	84,63	0,90
BRT metric	98,43	99,85	0,65

A second experiment, with the total number of records from PORBASE, was then used to test our similarity metric in a more realistic environment. In this case, we applied the exact matching and the BRT metric. For both cases, we measured the number of record matches and the corresponding number of FRBR works detected. The results are shown in Table 3.

Exact matching matched a total of 290.955 records, with 104.648 distinct works with an average of 2,78 records/work. Similarity matched a total of 355.840 records, forming 126.458 distinct works. It resulted in an increase of 22,3% in the number records matched and an increase in the number of groups of 20,8%, with an average of 2,82 records per work. The distribution of works by number of matched records on both methods can be seen in Table 4.

Table 3. Number of records matched and total sets of records created with exact matching and with the defined similarity metric in the PORBASE data set

	Exact matching	Similarity matching
Records matched	290.955	354.773
Works	104.648	126.457
Records per work (standard deviation)	2,78(2,86)	2,82(3.98)

Table 4. Distribution of works by number of records

Record per work	1	2	3	4	5	6	7-9	10+
Exact matching	1.069.731	73.980	16.276	6.059	2.899	1.668	2.077	1.689
Similarity matching	1.005.913	87.613	20.303	7.674	3.637	2.049	2.587	2.078

6 Discussion

Measuring the similarity between titles with a simple application of a generic similarity metric will result in good recall but low precision. On the other hand, exact matching results in good precision but low recall. The low recall obtained using exact matching on the Porto Editora data set was probably due to high number of records of school manual records that contained several subtitles, because the results obtained from the experiment with PORBASE revealed a smaller difference in the number of record matches between the exact and similarity methods.

Due to the lack of heterogeneity in author names in the Porto Editora dataset, it was not possible to evaluate the performance of similarity techniques for matching author names, and it was not possible to try to find a suitable data set due to lack of resources to manually check the record matches. However we don't believe that this

leads to the conclusion that similarity analysis of authors is of little use in the identification of work entities. It may still be relevant for records from unrelated sources. An example of such case is the LEAF project (Linking and Exploring Authority Files), which attempted to link authority records from libraries and archives from various European countries. This project, however, has only used exact matching for comparing person names [11].

For the above reasons, our experiment had a more focused approach on title similarity. The close observation of the mismatches that caused low recall in exact comparison and low precision in similarity metrics lead us to the development of a similarity metric specific for titles. The results obtained for both recall and precision were very close to 100%, leading us to conclude that it can be used in real world applications with significant increases in the usability of the library systems, with an insignificant introduction of errors by wrong record matching.

An interesting result was the high number of matches found by both matching methods. Matches by similarity were found in 354.773 records, representing 26% of the records in PORBASE.

7 Conclusions and Future Work

Our work has shown that similarity metrics can be used in the task of identifying FRBR works within bibliographic databases with a low error margin, while managing to identify most matches. When compared with exact matching the number of matches increases by a significant proportion. This proportion is very likely to be higher when trying to identify the works in more heterogenic environments, such as within libraries from different countries or in organizations of different types, archives and entertainment (theatre, cinema, etc.). We plan to further test our method in such environments. Likely points for improvement are in the matching of author names and on adding machine learning techniques to fine tune the final reasoning that matches the records.

The same techniques used in our work are likely to have applicability in other tasks related to FRBRization of bibliographic databases. These tasks include, for instance, identifying different expressions of the same work. It can also complement the work in [15] by identifying the duplicate entity instances extracted from the bibliographic records individually.

Our experience with PORBASE will also evolve to be integrated in a new prototype of an FRBR aware OPAC that is now under development.

References

1. IFLA Study Group on the Functional Requirements for Bibliographic Records: Functional requirements for bibliographic records: final report. München: K.G. Saur, UBCIM publications, new series, vol. 19 (1998), www.ifla.org/VII/s13/frbr/frbr.pdf ISBN 3-598-11382-X
2. Salton, G., Buckley, C.: Term-weighting approaches in automatic text retrieval. Information Processing & Management 24(5), 513–523 (1988)
3. Elmagarmid, A.K., Ipeirotis, P.G., Verykios, V.S.: Duplicate Record Detection: A Survey. IEEE Transactions on knowledge and data engineering 19(1), 1–16 (2007)

4. Bilenko, M., Mooney, R.J., Cohen, W.W., Ravikumar, P., Fienberg, S.E.: Adaptive name matching in information integration. IEEE Intelligent Systems 18(5), 16–23 (2003)
5. Zhao, M.: Semantic matching across heterogeneous data sources. Communications of the ACM 50(1), 45–50 (2007)
6. Zhao, H., Ram, S.: Entity identification for heterogeneous database integration: A multiple classifier system approach and empirical evaluation. Information Systems 30(2), 119–132 (2005)
7. Hickey, T.B., O'Neill, E.T., Toves, J.: Experiments with the IFLA Functional Requirements for Bibliographic Records (FRBR). D-Lib Magazine 8, 9 (2002), http://www.dlib.org/dlib/september02/hickey/09hickey.html
8. California Digital Library.: The Melvyl Recommender Project. Full Text Extension. Supplementary Report (2006), http://www.cdlib.org/inside/projects/melvyl_recommender/report_docs/mellon_extension.pdf
9. Cohen, W.W., Ravikumar, P., Fienberg, S.E.: A Comparison of String Distance Metrics for Name-Matching Tasks. American Association for Artificial Intelligence (2003), http://www.isi.edu/info-agents/workshops/ijcai03/papers/Cohen-p.pdf
10. Jaro, M.A.: Advances in record linking methodology as applied to the 1985 census of Tampa Florida. Journal of the American Statistical Society 64, 1183–1210 (1989)
11. Kaiser, M., Lieder, H.J., Majcen, K., Vallant, H.: New Ways of Sharing and Using Authority Information. D-Lib Magazine 9, 11 (2003), http://www.dlib.org/dlib/november03/lieder/11lieder.html
12. Lawrence, S., Giles, C.L., Bollacker, K.D.: Autonomous Citation Matching. In: Proceedings of the Third International Conference on Autonomous Agents, ACM press, New York (1999)
13. Pasula, H., Marthi, B., Milch, B., Russell, S., Shpitser, I.: Identity Uncertainty and Citation Matching. In: Advances in Neural Information Processing (2002), http://people.csail.mit.edu/milch/papers/nipsnewer.pdf
14. Lee, D., On, B.W., Kang, J., Park, S.: Effective and Scalable Solutions for Mixed and Split Citation Problems in Digital Libraries. In: Proceedings of the 2nd international workshop on Information quality in information systems, pp. 69–76 (2005)
15. Aalberg, T.: A process and tool for the conversion of MARC records to a normalized FRBR implementation. Digital Libraries: Achievements, Challenges and Opportunities. In: 9th International Conference on Asian Digital Libraries, pp. 283–292 (2006)