Signature Based Duplication Detection in Digital Libraries

L. Padmasree, Vamshi Ambati, J. Anand Chandulal, M. Sreenivasa Rao

Department of Electronics & Communication Engineering, VNRVJ Institute of Engg & Technology, Hyderabad, 072, India, lpsree@rediffmail.com
Regional Mega Scanning Center, International Institute of Information Technology, Hyderabad, India,vamshi@iiit.ac.in
Department of Computer Science, Vignan's Engineering College, Guntur, 522213, India. profanandlal@yahoo.co.in
School of Information Technology, JNT University, Hyderabad, 500 072, India srmeda@gmail.com

Abstract -- **Duplications should be removed to improve both efficiency and effectiveness of an information Retrieval system. In Digital Libraries due to varied sources of books that are distributed across various parts of the country, duplicates could arise between scanning points. The Duplication of the books can be identified only using metadata of a book. If the metadata is incorrect, abbreviated, missing or incomplete it makes the duplicate detection all the more difficult. This paper discusses a technique that works fast and efficiently in detecting the duplication of the books. Duplicate detection was done by similarity search using signature file method where we can detect the duplicate with typographical mistakes, word disorder, inconsistent abbreviations and even with missing words. The performance of the similarity search is efficient since all the signatures are in the binary format and computations are done by low level logical operations.**

Index Terms – Metadata, Similarity search, Signature

I. INTRODUCTION

Digital Libraries have received wide attention in the recent years allowing access to digital information from anywhere across the world [1][2]. Traditionally, digital libraries work in a closed environment and contain the process of information and the content in a local repository. Although doing so increases the ease of server management and administration. Such an isolated set up does not scale up easily or promote collaboration across geographically distributed points of operation, for which a distributed environment becomes a requisite where the entire workflow can be automated. This raises certain operational and policy related problems and challenges such as procurement of books, incomplete and incorrect metadata, duplication and data management. Most of the books scanned in a digital library are procured from various sources that are distributed across various parts of the country, duplicates could arise between scanning points. The Duplication of the books can be identified only using metadata (title, author, publishing year, edition, etc) of a book. However, if the metadata is incorrect, missing or incomplete it makes the duplicate detection more difficult. This paper discusses a technique that works fast and efficient in detecting the duplication of the books.

Duplicate detection was done by similarity search using signature file method where we can detect the duplicate with typographical mistakes, word disorder, inconsistent abbreviations and even with missing words. The above technique is applied on the metadata of Digital Library of India 'DLI'[3] and the results of the duplication detection are depicted in Table III which support the above statement.

In this paper the section II deals with motivation and section III gives related work. The implementation details are presented in section IV and the section V contains the experimental results.

II. MOTIVATION

Most of the books scanned in the DLI project are procured from sources like libraries and government archives and hence already contain metadata entered by knowledgeable personnel, which can be relied upon, but is still debatable due to individual biases. However a major portion of the sources of books in the project have metadata only in non-digital formats and so these have to be fed into the system manually. This process though inevitable is understood to be prone to errors. Due to these varied sources of book flow in the DLI project in multiple languages and due to the lack of standard formats, metadata is missing, incorrect or incomplete or sometimes difficult to interpret. Inaccurate metadata hinders fruitful search and retrieval of books, categorization and at the same time most importantly brings in scope for duplicate entries of the same book. Duplicates could arise between

scanning locations maintained by the DLI project. Effort put into scanning a book,[4]processing the images and quality assurance can not be afforded to be spent on duplicates. Communicating metadata across centers and within scanning locations is important. The Duplication of the books can be identified only using metadata of a book like the title, author, publishing year, edition, etc. However, if the metadata is incorrect, missing or incomplete as discussed in the next section, it makes the duplicate detection all the more difficult.

In order to tackle this issue of duplicates, at DLI prior to scanning the metadata of books is first uploaded to a central repository. At this repository duplication check takes place. This process needs to be as quick as possible so that the scanning of the books does not have to be delayed at the scanning locations. A direct comparison of title and author with the existing ones may not be a problem. But given the discrepancies in the metadata of the books like spell errors, jumbled words, naming convention variations the problem turns out to be more interesting. Therefore there is an immense need for such effective and efficient duplicate detection algorithms.

III. RELATED WORK

Most of the methods for detecting duplicates depend on finding similarity between documents [5]-[7]. Most traditional methods for calculating string similarity can be roughly separated into two groups: character-based techniques and vector space based techniques. The former rely on character edit operations, such as deletions, insertions, substitutions and subsequence comparison, while the latter transform strings into vector representation on which similarity computations are conducted. While character-based metrics work well for estimating distance between strings that differ due to typographical errors or abbreviations, they become computationally expensive and less accurate for larger strings [8].

The vector-space model of text avoids this problem by viewing strings as “bags of tokens” and disregarding the order in which the tokens occur in the strings. Given a vocabulary of n possible tokens in a corpus, strings are represented as sparse n-dimensional vectors of real numbers, where every nonzero component corresponds to a token present in a string. Researchers have examined several metrics for determining the similarity of a document to another document. TF-IDF is the most popular method for computing the weights of the vector representation; it takes into account token frequencies within a particular string and over the entire corpus [8].

In the present work we used an efficient and fast duplication detection technique using similarity search. Here duplicate records can be detected not only exact match but also approximate matching due to spell mistakes, missing words and jumbled words. Basically we use a Signature file approach. The signature file approach seems most promising for large data base as it has good text retrieval properties and require small storage over head[9].To detect the duplicates in the metadata of the library is as follows: A Signature is computed for the meta data of each book in the library database. The method of computing the signature by using hashing and superimposed coding techniques is discussed in section IV. The signatures of the metadata of all the books are stored as a signature file. The metadata of the new book, which is going to be scanned, is used as a query. The Similarity search algorithm retrieves the approximate duplicate of it, if it exists. The Similarity search algorithm finds the Jaccard distance to detect the approximate duplicate record. As the signatures are stored in binary form the approximate duplicate detection is fast, efficient and storage effective. The proposed technique provides another important feature that supports language independency. It is possible since the signatures are created from ASCII format and the other languages metadata, apart from English can also be generated in the ASCII format by using tools like OM Transliteration Tool.

IV. IMPLEMENTATION

This section deals with the implementation of Duplicate Detection in Digital Library databases. The figure 1. gives the broad process of duplicate detection in Digital libraries. As the books are scanned their metadata which is non-digital formats is fed into the central repository manually. This metadata is converted into a binary signature and stored in the signature file.

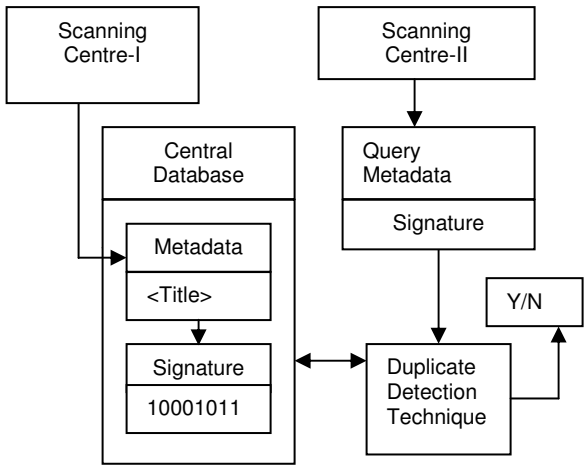

Fig.1. Duplicate Detection in Digital Library system

TABLE II

Central Repository	
Metadata of Books	Signatures
The Meaning And Teaching Of Music -Will Earhart	01111111000010111110001111101 1
Some Famous Singers Of The 19th Century -Francis Rogers	11100101000000100111111010110 110
A Dictionary of Musical Terms - Dr.th.baker	11110010100011010000011111111 1
The Arts of Japan - Edward Dillon	11110110000000000000000110011 11

Query - Spell Mistakes	Query - Missing Words	Query - Jumbled Words
The Arts of Japa Edward Dillon	The of Japan - Edward Dillon	Edward Dillon -The Arts of Japan
11110110000111000000011001111	11110110000001100000011001111	11110110000010000000011001111

Result : The Arts of Japan - Edward Dillon

When a new book is going to be scanned from another scanning center its metadata is used as a query to detect whether its duplicate exists in the central repository or not. However, if the metadata is incorrect, missing, incomplete or disorder the exact string is not possible to detect duplicates directly. To have a close proximity match this query is converted to a signature using same encoding technique that is used for forming record signature. An effective way to form the signature is by using superimposed coding technique [10].

A. *Super Imposed Coding Technique*

In Superimposed Coding Technique each record is mapped into an individual binary signature. Record is either the title or the author name of the book or the combination. Signatures of the records in the training data and testing data are encoded binary representations, which characterize the essence of them. Here, the signature of the 'title or author name' of the book is obtained by superimposing the signatures of the words with OR operation. The signature of each word is obtained by hashing technique discussed in the next section 4.B. Now the computational steps involved in the superimposed coding technique are presented an example as follows.
Example: If a book with title "Computer Programming" consists of two words, 1.Computer 2. Programming. Let the signature be an n-bit pattern, in which r-bits are set to 1. Then it is one among the ${}^{n}C_{r}$ bit patterns that can be generated using n bits in which r bits are set to 1. If n = 12 and r = 4 the signatures of the words are shown in table I. The signature of the title of the book is obtained by superimposing the signatures of the words with OR operation.

TABLE I

Computer Programming	1100 0001	1000 0101	0100 0100
Signature of the book	1101	1101	0100

B. *The Algorithm for Hashing Method*

In order to get the signature of a word, hashing technique is used[11]. The hashing function H(w) maps the word(w)into one of the patterns generated by computing a hash value of the word. The hash function uses shift and add strategy. The ASCII values of the characters in the word are added and shifted by H(w) in order to compute the hash value. The final hash value is obtained by mod operation with ${}^{n}C_{r}$.
Example: The signature for the word "COMPUTER" is calculated as follows: The table size is ${}^{n}C_{r}$, here n = 4, r = 2. So table size = ${}^{4}C_{2}$ =6. The Hash value of "COMPUTER" = 18612 mod 6 = 0. So the signature of the word "COMPUTER" = 0011.

C. Duplicate Detection in Digital Library System

The duplicate detection in digital library uses the signature file method. In the signature file method a signature or descriptor is associated with each metadata of the scanned book in the database. The signature is a bit encoding of the values used to retrieve the record. The similarity search using the jaccard distance. i.e., the document is retrieved whose signature is with minimum distance with that of the query. The computational steps involved in the library database system are given as an algorithm and is presented in Fig.2. The following example illustrates the process of duplicate detection for digital library database.

Input **:** L library database consists of documents D1, D2, ……, Dm, Q query.
Output : B book corresponding to query Q
Procedure Library (D1, D2, ……,Dm, Q : in; B : out)

1. *for i=1 to m do*
2. Si = superimposed-coding(Di)
3. *end do*
4. X = superimposed-coding(Q)
5. O = Jaccard (S1, S2,……Sm, X)
6. Look up in Library database L for a book B (document) whose Signature matches with minimum Jaccard distance.
7. End

Fig. 2. The Similarity Match Algorithm for Library Database

Example: Let us consider a sample library database that consists of four books. Each book in the library database has the title and author name. Details of four books are shown in table II.
The query with spell mistakes, missing words and jumbled words for the book 'The Arts of Japan - Edward Dillon" is retrieved since it has minimum jaccard distance from all other signatures.

D. Jaccard Distance

The Jaccard distance between the query signature and target signature can be obtained by using the expression as $d = (r + s) / (q + r + s+t)$ where q is the number of variables that equal 1 for both target and query signatures, r is the number of variables that equal 1 for target signature but that are 0 for the query signatures , s is the number of variables that equal 0 for the target signature but equal 1 for the query signature , and t is the number of variables that equal 0 for both target and query signatures .

V. EXPERIMENTAL RESULTS

The performance of the duplicate detection of the Digital Library system critically depends on the efficiency of signature computation and the efficiency of signature computation is in turn depends on the appropriate choice of two parameters n and r. In the event of occurrence of false drop, it may happen that two distinct records may correspond to same signature. It is observed that for appropriate values of n and r, false drop can be avoided. Experiments were carried out to evaluate the performance of system in the context of Digital Library of India (DLI). As false drops may increase with the increase in size of the metadata repository, our aim is to study the scalability and accuracy of the system under this condition as shown in fig 3.

Experiments were conducted by varying the size of the signature and fixed at 75 as it is giving efficient retrieval rate. Experiment was carried out with 15% of input records taken as query records which differ from the input records in the following three cases. The first case deals with the spelling mistakes in the input record. The second case deals with the case in which one or more words differ (even deleted) from the input record. The third case deals with the jumbled words in the input records. The results are shown in table III.

TABLE III

Meta data	Query-Spell mistakes		Query-Missing Words		Query-Jumbled Words	
	false drop (%)	DR (%)	false drop (%)	DR (%)	false drop (%)	DR (%)
1000	7	93	9	91	3	97
5000	8	92	10	90	5	95
23000	10	90	12	88	5	95

DR= Detection Rate

Based on the observations, it is concluded that on an average percentage of duplicate detection by the system is around 92. The percentage of duplicate detection is defined as the product of the ratio of the number of accurate retrievals to the number of input queries and 100. Still the false drops can reduced by clustering the library data according to category and improve the duplicate detection rate.

For instance, consider the grouping of the metadata of 100,000 books into 10 categories like chemistry,

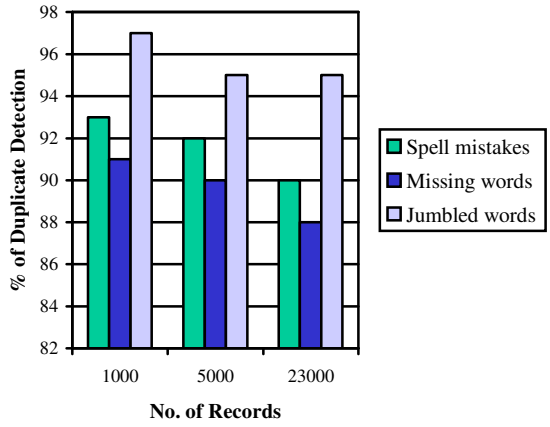

Fig .3. Scalbility and accuracy of duplicate detection system

biology and so on. When a query string consisting of biology category is submitted to the system, the signature mapping takes place within the biology category rather than the entire 100,000 books. This reduces the false drops since the signatures are computed on a small range of books.

The problem of false drops misleads the user in identifying the appropriate duplicates. If the user enters either the title or the author name for searching duplicates, then it might cause confusion, so when applying this similarity search technique, we require the user to enter the title, the author and the similarity distance.

This helps in listing the range (instead of the one that closely matches) of combination of titles and authors that are in nearest match. This range falls into small close similar categories rather than the entire collection of books as shown in fig.4. However, if proper signature computation scheme is employed the retrieval performance is believed to be improved.

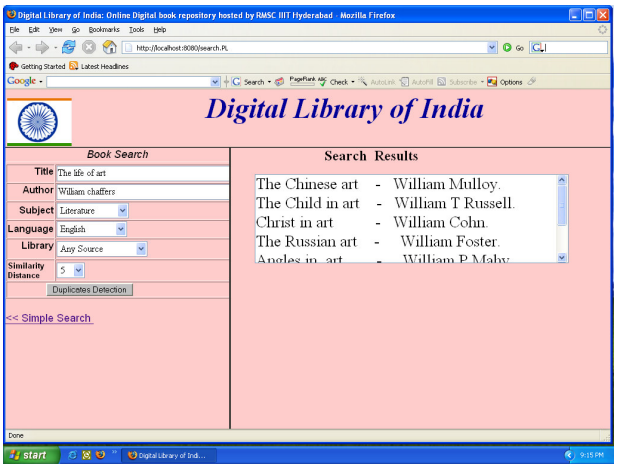

Fig 4.

VI. CONCLUSION

In this paper an effective and efficient duplicate detection technique is proposed. Duplicate detection was done by similarity search using signature file method where we can detect the duplicate with topographical mistakes, word disorder, and inconsistent abbreviations and even with missing words. The performance of the technique is examined in the duplicate detection process of DLI and shown that the percentage duplicate detection is around 95. Even this decrease in the performance of the duplicate detection can be avoided by another appropriate signature method. Another highlight is that it provides language independency.

ACKNOWLEDGEMENT

We owe our sincere thanks to Prof. Raj Reddy, Carnegie Mellon University, Pittsburgh, USA and the contribution made towards by the Digital Library teams at International Institute of Information Technology, Hyderabad, India. We also acknowledge the efforts of the reviewers and their feedbacks on this paper.

REFERENCES

[1] Michael Lesk, Understanding Digital Libraries, Morgan Kaufmann, 2004.

[2] Alexa T.McCray., Marie E. Gallagher." Principles for Digital Library development", *Communications of the ACM* , 2001.

[3] Vamshi Ambati, N.Balakrishnan, Raj Reddy, Lakshmi Pratha, C V Jawahar: The Digital Library of India Project: Process, Policies and Architecture, *In the Proceedings of 2nd International Conference on Digital Libraries(ICDL)*, 2006.

[4] Vamshi Ambati, Pramod Sankar, Lakshmi Pratha, C.V.Jawahar " Quality management in digital libraries"

[5] Chowdhury, A ., Frieder . o ., Grossman . D. ,and McCabe. M. C. "Collection statics for fast duplicate document detection".*ACM Transaction on Information Systems*, 2002.

[6] M.Bilenko and R.J.Mooney."Learning to combine *trained distance metrics for duplicate detection in* databases".Technical Report AI 02 – 296, Artificial Intelligence Laboratory , University of Texas at Austin , Austin TX Feb,2002.

[7] Cho.J. ,Shivakumar . N ., and Garcia –Molina . H. Finding replicated web collections*".In proceedings of the ACM SIGMOID Conference on Management of Data* ,1999.

[8] Mikhail Bilenko and Raymond J. Mooney, "Adaptive Duplicate Detection Using Learnable String Similarity Measures", *Proceedings of the Ninth ACM SIGKDD International Conference* on Knowledge Discovery and Data Mining, 2003.

[9] Sacks –Dany, R ,Kent .A, Ram Mohan Rao.K "Multikey Access Methods Based on

Superimposed Coding Technique". *CM Transactioins on Database System* ,1987..

[10] Faloutsos.C."Access methods for text" *, ACM Computing Surveys.*1985,

[11] Sreenivasa Rao, M., Pujari, A. K., Sreenivasan, B. "A new neural network architecture for efficient close proximity match of large databases*"*. *IEEE Computer Society Press, Proceedings of the Eighth International Workshop on DEXA, France, Edited by R. R. Wanger*, 444-449, 1997.

[12] S. B. Needleman and C.D. Wunch."A general method applicable to the search for similarities in the amino acid sequences of two proteins.Journal of Molecular Biology ", 1970.

[13] Shang ,H. ,Merrettal ,T. H.,"Tries for Approximate String Matching knowledge*", IEEE trans on ge and data Engineering* ,1996.

[14] Bethina Schmitt and Sven berländer,"Evaluating and Enhancing Meta-Search Performance in Digital Libraries.