INNOVATIVE TECHNOLOGICA
METHODICAL RESEARCH JOURNAL
ISSN: 2776-0987 Volume 3, Issue 12 Dec. 2022

LINKING MODEL OF BIBLIOGRAPHICAL DATABASE

Ishniyazov O.O.
Tashkent University of Information Technologies
named after Muhammad al-Khwarizmi, Senior Teacher,
oishniyazov@gmail.com

Babajanov M. R.
Tashkent University of Information Technologies
named after Muhammad al-Khwarizmi, PhD
mominbabajanov@gmail.com)

Shokirov Sh.Sh.
Tashkent University of Information Technologies
named after Muhammad al-Khwarizmi, Senior Teacher,
shodmonimomov@gmail.com

Abstract:
This text devoted to model of bibliographical database. It is also mentioned about the functional blocks of preparation, creation of pair records, comparison of fields with pair Records, decision-making, which constitute the algorithm.

Keywords: bibliographic record, electronic catalog, information-library systems, information-resource centers, full text, Electronic Library, automated information-library systems, bibliographic description.

Introduction

In the development of information and communication technologies around the world, the issues of identification of bibliographic information of objects, the constant growth of information volumes, as well as the development of scientific research are becoming increasingly important. The rapidly growing amount of information available to a wide range of users in the evolving digital world is causing some problems in the search process. It is essential for storing, organizing, and organizing this information. Because when we search for information, we come across information that has a lot of similarities.

ISSN: 2776-0987 Volume 3, Issue 12 Dec. 2022

Therefore, it is permissible to identify such data as duplicate data and present them as one of the output data.

Description of the current state

The need to identify real-world objects in bibliographic data was considered in the late 19th century by Paul Otle, a Belgian sociologist, documentary scientist, bibliographer, lawyer, and one of the founders of computer science theory [1, 2]. Such identification can be done by contacting a special record that clearly indicates this object. Such a record may be any structured document containing information about the object and meeting the requirements developed by international organizations.

Currently, computing systems of various publications, such as Scopus, Web of Science, SCIENCE INDEX (based on RSCI), use identification codes of different authors.

However, a single author can be registered in different databases with different codes. Thus, linking these codes to each other is a topical issue [3]. It is also possible to link the bibliographic records used in information resource centers to a database name or a field value.

The development of this approach is able to improve the quality of scientific indicators, taking into account the authorial publications recorded in various databases.

The process of establishing a link between authoritarian and bibliographic records within the existing Automated Information Library Systems is carried out by the cataloger. On the one hand, an expert can establish a sufficiently reliable relationship between records by attracting additional information that is not in the records themselves. On the other hand, such an approach involves a large amount of manual labor, the complexity of retrospective analysis, and many “lost” connections between records.

Development of the mathematic model

The data search process is based on bibliographic descriptive elements. The data entered for the search are compared with the entries in the bibliographic database, and the information that belongs to an individual and to which the elements of the bibliographic description are appropriate is given as a result [5].

INNOVATIVE TECHNOLOGICA
METHODICAL RESEARCH JOURNAL
ISSN: 2776-0987 Volume 3, Issue 12 Dec. 2022

The record linking model is based on specific functional requirements for bibliographic records. The publisher, author, and organization are treated as separate objects, and information about those objects is stored in parameter values that are components of the record. The compatibility of the records means that they represent the same object in the real world, and the records can only be compared with a set of parameter values.

Thus, let two sets of records A and B be given, - α(a) is a record in set A that describes an object; β(b) is a record that represents object b in set B. Objects can also belong to a set that is common. For example, information about authors around the world. a ∈ C, $b \in C \in$.

We define a set of pairs of records that describe one or more A objects in the real world:

$$S(\mathrm{a}) =< \alpha(\mathrm{a}), \beta(\mathrm{b}) > \qquad (0.1)$$

Where, $a = b$, $\alpha(\mathrm{a}) \in \mathrm{A}$ and $\beta(\mathrm{b}) \in \mathrm{B}$. This means that when we create a pair of records, we select similar records from different sets and create another set. If we combine the sets built for each object into a set C, we have a set S with a pair of records.

$D(\mathrm{a})$ represents a pair of records $S(\mathrm{a})$ describing different objects in a collection:

$$D(\mathrm{a}) =< \alpha(\mathrm{a}), \beta(\mathrm{b}) > \qquad (0.2)$$

Where $a \neq b$ and $\alpha(\mathrm{a}) \in \mathrm{A}$; $\beta(\mathrm{b}) \in \mathrm{B}$.

Similarly, by combining $D(\mathrm{a})$ for all possible objects, we obtain several incompatible pairs with the records of the set D.

Since the structure of records α and β may be different, rules must be developed to organize multiple pairs of records. We define the rules for comparing records as c_j. $j = \overline{1,K}$, each rule is generated by comparing the comparison function and the intersection of authority and bibliographic record fields.

The set of results applied to all K rules can be expressed as a point in space of the properties of K dimensions, i.e. $\gamma = (X_1,\ldots, X_k)^T$ is the result of applying the rule of comparison X_j - c_j here, which is the evaluation of the conformity of certain data in the records. This assessment is expressed in predefined gradations (degrees) that may be different for different rules.

INNOVATIVE TECHNOLOGICA
METHODICAL RESEARCH JOURNAL
ISSN: 2776-0987 Volume 3, Issue 12 Dec. 2022

To solve the problem of linking records, it is necessary to create a decisive function that serves to assess the authenticity status of objects based on the set of precedents available.

$$O(\gamma)=\begin{cases}1,\langle\alpha(\mathrm{a}),\beta(\mathrm{b})\rangle\in S,\\ 0,\langle\alpha(a),\beta(b)\rangle\in D,\end{cases} \tag{0.3}$$

Precedents are $\langle\alpha(a),\beta(b)\rangle$ pairs with a definite $L(a,b)$ status, from which a teaching pattern is formed.

$$L(a,b)=\begin{cases}1,\mathrm{a}=\mathrm{b},\\ 0,\mathrm{a}\neq\mathrm{b},\end{cases} \tag{0.4}$$

Let us present the instructional example as a set of two non-intersecting points in the character space. The first set summarizes a pair of similar records and represents a single object:

$$C^S=\left\{\gamma\middle|\langle\alpha(a),\beta(\mathrm{b})\rangle\in S\right\} \tag{0.5}$$

The second set consists of pairs depicting different objects

$$C^D=\left\{\gamma\middle|\langle\alpha(a),\beta(\mathrm{b})\rangle\in D\right\}$$

In this case, it can be determined whether the new record pair belongs to sets S and D by calculating the distance sought between sets C^S and C^D for classification. The choice of distance depends on the problem solving requirements. In this work, it is proposed to use the Mahalanobis distance method to determine the distance sought, taking into account the interdependence of the characters and the invariance of the scale.

The distance C to the center of the set the square of the distance Mahalonobis is calculated by the following formula:

$$Dist^2(\gamma,\mu^S)=(\gamma-\mu^S)W^{-1}(\gamma-\mu^S)^T \tag{0.6}$$

Where:

γ - the vector of the character values represents the result of applying one of the rules of comparison of each character, $\gamma=(X_1,...,X_k)^T$;

μ^S - S set center;

W^{-1} - is a covariance matrix that forms the inverse of the matrix;

D - the distance to the center of the set is the same.

$$Dist^2(\gamma,\mu^U)=(\gamma-\mu^U)W^{-1}(\gamma-\mu^U)^T \tag{0.7}$$

INNOVATIVE TECHNOLOGICA
METHODICAL RESEARCH JOURNAL
ISSN: 2776-0987 Volume 3, Issue 12 Dec. 2022

Here μ^D - center distance of the set of U.

The center is the arithmetic mean vector of the symbols, the components of which are calculated according to the following formula:

$$\mu_i^S = \frac{1}{n^S}\sum_{k=1}^{n^S} X_{ik}^S \qquad (0.8)$$

Where:

μ_i^S - μ^S - i companent of the vector

X_{ik}^S - γ_k - value of the i component of the vector, $\gamma_k \in C^S$, $k = \overline{1, n^S}$

The elements of the W covariance matrix are calculated as follows.

$$W_{ij} = \frac{1}{n^S + n^D - 2}\left\{\sum_{k=1}^{n^S}\left(X_{ik}^S - \mu_i^S\right)\left(X_{jk}^S - \mu_j^S\right) + \sum_{k=1}^{n^D}\left(X_{ik}^D - \mu_i^D\right)\left(X_{jk}^D - \mu_j^D\right)\right\} \quad (0.9)$$

Bu yerda:

n^S - the number of pairs taken from the instruction sample in the set C^S;

n^D - the number of pairs taken from the teacher sample in the set C^D;

X_{ik}^S - the value of the i-component of the character value vector for k-pair records in the set; (the size of the i-component of the character value vector) C^S;

X_{ik}^D - the i-component value of the character value vector for k-pair records in the set C^D;

μ_i^S - the average value of the i-component of the character value vector in the class C^S;

μ_i^D - the average value of the i-component of the character value vector in the class C^D;

The elements of the W^{-1} matrix calculated on the basis of the instructional pattern can be considered as important factors that reflect the importance of a particular character (or the specific rules of comparison of records). It is also possible to estimate the distance between the centers of the two sets of Mahalanobis in order of the proportion of characters, which makes it possible to distinguish the most important rules for communication.

As a criterion for creating an important function, it is possible to minimize the number of classification errors for a pair from the teacher sample.

INNOVATIVE TECHNOLOGICA
METHODICAL RESEARCH JOURNAL
ISSN: 2776-0987 Volume 3, Issue 12 Dec. 2022

$$\min_{\{c_j\}} \sum_{i=1}^{N} I\{O(\gamma_i) \neq L(a,b)\} \quad (0.10)$$

Where:

$\{c_j\}$ - is a set of rules to compare and decide on whether a record is appropriate or not, $j = \overline{1,K}$;

I - identificatory function;

γ_i is the sign value vector for i-pairs of records in the test result, $i = \overline{1,N}$, and N is the number of pairs of records in the test sample;

Thus, the number of errors in the different sets of comparison rules used to compare records is minimized. Obviously, different rules lead to different mistakes. After checking the quality of the classification based on the test sample, you should select a set that allows you to achieve fewer errors.

Conclusion

Resolving this issue will automatically provide information that belongs to the author as a result. Articles written by our author are currently available in the articles database, and there are some shortcomings when searching for articles related to the author. That is, these shortcomings were considered as follows:

1. In some cases, a bibliographic description of the author's articles is not provided. This is of course due to some changes in registration (for example, full names are written differently in different languages);

2. In order to eliminate the first shortcoming, of course, we need to contact the system administrator and leave a message about the articles belonging to the author. Once the system has been reviewed by the organizers, the data per author will be combined.

The intended part of the program serves to eliminate the above-mentioned shortcomings and allows you to automatically link the author's bibliographic information.

References

[1] Отле П. Библиотека, библиография, документация [Текст] : Избранные труды пионера информатики / ПольОтле. – Москва: ФАИР-ПРЕСС: Пашков Дом, 2004. – 348, [1] с.

ISSN: 2776-0987 Volume 3, Issue 12 Dec. 2022

[2] Отле П. Труды по библиотековедению. Руководство для общественных библиотек. Организация умственного труда. Руководство к администрированию [Текст] : Практ. пособие / Поль Отле; [Вступ. ст. и науч. ред. Ю. Н. Столярова]. – Москва : Либерея, 2002. – 227 с. : табл. – ISBN 5-85129-148-6.
[3] Мазов Н.А. Новые методы формирования публикационного профиля научной организации в сети науки / Н. А. Мазов, В. Н. Гуреев // Науч. И техн. б-ки. – 2013. - № 12. – С. 42-48.
[4] Manning C. D. Introduction to Information Retrieval [Electronic resource] / C. D. Manning, P. Raghavan, H. SchuËtze – Cambridge, 2009–2011.
[5] Winkler W. E. Overview of record linkage and current research directions [Electronic resource] : tech. report / W. E. Winkler ; U.S. Census Bureau, Stat. res. div. – Washington : [s. n.], 2006. – 44 p. – (RRS (Statistics #2006-2)). – URL : http://www.census.gov/srd/papers/pdf/rrs2006-02.pdf, free. – Tit. From the screen (usage date: 04.06.2013).
[6] Talburt J. Entity resolution and information quality / John R. Talburt. – San Francisco : Morgan Kaufmann/Elsevier, 2011. – 256 p.
[7] Needleman S. B. A general method applicable to the search for similarities in the amino acid sequences of two proteins / S. B. Needleman, C. D. Wunsch // J. mol. biol. – 1970. – Vol. 48, № 3. – P. 443–453.
[8] Monge A. E. The field matching problem: Algorithms and applications / A. E. Monge, C. P. Elkan // Proc. 2nd Int. conf. on knowledge discovery and data mining (KDD-96), Portland, OR, USA, Aug 2–4, 1996. – Portland : AAAI Press, 1996. – P. 267–270.
[9] Cilibrasi R. Clustering by compression / R. Cilibrasi, P. M. B. Vitanyi // IEEE. trans. on inf. theory – 2005. – Vol. 51, № 4. – P. 1523–1545.
[10] Elfeky M. G. TAILOR: a record linkage tool box / M. G. Elfeky, A. K. Elmagarmid, V. S. Verykios // Proc. 18th Int. conf. on data eng. (ICDE 02), San Jose, CA, USA, 26 Febr.–1 March, 2002. – Washington : IEEE Computer Soc., 2002. – P. 17–28.