Выявление дубликатов в библиографических базах данных

© А.А. Князева © И.Ю. Турчановский © О.С. Колобов
Томский филиал Института вычислительных технологий СО РАН Томск
aknjazeva@ict.nsc.ru tur@hcei.tsc.ru
Институт сильноточной электроники СО РАН Томск
okolobov@hcei.tsc.ru

Аннотация

В работе рассматривается задача выявления дублирующихся записей в электронном каталоге библиотеки. Предлагается модель выявления дубликатов, основанная на обучении с учителем. Обучающая выборка, позволяющая настроиться на особенности конкретных баз данных, строится на основе тех записей, для которых известен идентификатор ISBN или ISSN. Далее вычисленные на основе обучающей выборки весовые коэффициенты используются для работы с записями, в которых отсутствуют идентификаторы ISBN и ISSN.

1 Введение

В современных библиотечных системах важной проблемой является выявление дублирующихся записей. Рост интереса к этой проблеме обусловлен, с одной стороны, увеличением объемов информации, подлежащей хранению, а с другой стороны – развивающейся интеграцией информационных ресурсов. В процессе объединения электронных каталогов различных библиотек необходимо учитывать возможные повторения информации, когда одна и та же публикация описывается несколькими библиографическими записями. При этом библиографические записи, описывающие одну и ту же публикацию могут существенно отличаться друг от друга. В частности, такие записи могут различаться по своей структуре, иметь различную степень полноты, содержать перестановки слов, опечатки и т. д. [1].

Наличие дублирующихся записей приводит к необходимости объединения их в единую запись или удаления избыточных записей. Для этого необходим механизм сопоставления записей, позволяющий сделать вывод об их соответствии или несоответствии друг другу. Соответствие записей означает, что в них упоминается один и тот же объект реального мира. Таким образом, речь идет об идентификации этого объекта на основе сведений, содержащихся в записях.

Труды 15-й Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» — RCDL-2013, Переславль-Залесский, Россия, 14–17 октября 2013 г.

2 Связанные работы

Выявление дубликатов является достаточной сложной и комплексной задачей. Эта задача, в свою очередь, включает в себя следующие:

- предварительная подготовка записей, в которую при необходимости может входить проверка на корректность значений отдельных полей, сопоставление этих значений с используемыми словарями и т.п.;
- способ составления пар из записей, который позволяет сократить количество рассматриваемых пар;
- способы сравнения отдельных полей записей, позволяющие учесть нечеткое соответствие их значений;
- принятие решения о соответствии на уровне записи на основе результатов сопоставления отдельных полей.

Системы выявления дубликатов различаются между собой способом решения перечисленных задач [2]. В центре внимания данной работы находится задача принятия решения о соответствии. В литературе ее не всегда выделяют, зачастую ограничиваясь перечислением методов сопоставления отдельных полей. В то же время, набор оценок соответствия по всем признакам необходимо агрегировать в единую оценку соответствия записей. При этом одни признаки могут нести больше информации, чем другие. Задача присвоения признакам соответствующих весовых коэффициентов является нетривиальной и не всегда ее можно решить с помощью эмпирических правил.

Диапазон существующих систем выявления дубликатов достаточно широк: техники для установления RDF-ссылок в Веб, базы данных с адресами

клиентов и организаций, системы для связывания демографической и медицинской информации о персонах и системы, работающие с библиографической информацией.

В отдельную группу можно выделить методы, предназначенные для связывания данных в Веб с помощью установления RDF-ссылок (Silk [3], Oyster Entity Resolution [4]). Хотя в основе таких методик лежат те же общие предположения, что и во всех системах связывания, большой отпечаток на них накладывает специфичность конкретных данных. Это не позволяет непосредственно использовать данные системы для решения рассматриваемой задачи.

Наиболее многочисленной является группа систем, настроенная на поиск дубликатов в одном или нескольких текстовых файлах (либо в реляционых базах данных), содержащих сведения об именах, почтовых адресах, телефонах, номерах страховки и т.п. Чаще всего, для принятия решения о соответствии записей используется набор правил или классическая вероятностная модель F-S. В первом случае система предоставляет пользователю возможность определения того, насколько важно совпадение по тому или иному признаку (например, системы Silk и Oyster Entity Resolution). На практике это может быть достаточно трудно сделать, поскольку не всегда в распоряжении пользователя есть такая информация. Во втором случае вес того или иного признака вычисляется автоматически, на основании функции правдоподобия (например, система AutoMatch [5]). При этом принимается предположение о функции распределения признаков, а также их взаимной независимости. Также, многие системы предлагают использование одного из описанных механизмов на выбор пользователя (Febrl [6], FRIL [7], Merge Tool Box [8]). Системы этой группы отличаются друг от друга гибкостью настройки, инструментами для нормализации данных и сравнения отдельных полей, возможностями визуалицации и т.п. К недостаткам систем этой группы с точки зрения решаемой задачи можно отнести то, что они не поддерживают данные сложной структуры, требуют установления правил связывания в явном виде или принятия предположений о функции распределения признаков и их взаимной независимости, которые не обяательно выполняются на практике.

Группу систем, работающих с библиографическими данными, в свою очередь можно разделить на две части: системы для «простого» формата библиографической ссылки (такого как BiBTEX или неструктурированная библиографическая запись) и системы для работы с «профессиональными» форматами (семейство MARC-форматов). Первая группа систем вынуждена больше внимания уделять такой частной задаче, как автоматическая разметка неразмеченного текста (чтобы выделять из текстовой строки элементы библиографического описания). Во второй группе такая необходимость отпадает благодаря сложной структуре формата, но в то же время появляется необходимость учета этой структуры, в которой одна и та же информация может быть внесена поразному, в зависимости от предпочтений каталогизаторов. К первой группе можно отнести системы DIFWICS [9], MARLIN [10] и механизм, используемый в системе Mendeley[1] [11] (поскольку система Mendeley нацелена на работу с полнотекстовыми записями, одним из принципов выявления дубликатов в ней является сравнение хэшей файлов). Ко второй группе относятся проект VIAF [12, 13], система, используемая консорциумом SC LENDS[2], и представляемая в данной работе модель выявления дубликатов. Однако хотя в проекте VIAF и реализована работа с данными в MARC-формате, он не предлагает механизма для автоматической оценки весов признаков, поскольку основан на использовании эмпирических правил. Таким образом, несмотря на некоторое сходство, не представляется возможным заимствовать подход, использованный в проекте VIAF для решения поставленной задачи. Система, разработанная для консорциума SC LENDS [14], основана на сопоставлении по двум параметрам: заглавие и ISBN[3] или ISSN[4]. В данной работе мы исходили из того, что библиографические записи не всегда содержат корректные идентификаторы ISBN/ ISSN, поэтому и данный подход не может быть применен для решения поставленной задачи.

В рамках данной работы существующие системы идентификации объектов рассматривались с точки зрения следующих критериев:

1. Отсутствие предположений о функции распределения признаков;
2. Отказ от эмпирических правил для принятия решения о соответствии записей;
3. Отсутствие требования независимости сравнительных признаков;
4. Работа с записями в форматах семейства MARC;
5. Возможность работы с неполными данными, в том числе и записями, в которых не указаны идентификаторы ISBN/ISSN.

Как показал проведенный анализ, в данный момент не существует системы, которая отвечала бы всем приведенным требованиям.

---

[1] http://www.mendeley.com/

[2] South Carolina Library Evergreen Network Delivery System (англ.)

[3] International Standard Book Number (англ.) – Международный стандартный книжный номер

[4] International Standard Serial Number (англ.) – Международный стандартный серийный номер

3 Предлагаемый подход

Ранее, в наших работах [15,16] был предложен механизм идентификации объектов реального мира, упоминаемых в структурированных записях различных типов. Очевидно, задачу выявления дубликатов можно рассматривать как частный случай данной задачи, в котором анализируются записи одного типа.

Суть предлагаемого подхода заключается в том, чтобы обучать функцию принятия решения о соответствии записей на основе обучающей выборки. Такая выборка должна состоять из пар записей, для которых известен истинный статус соответствия (т.е. относятся они к одной публикации или к разным). Следует задать набор признаков для сопоставления записей и вычислить значения этих признаков для каждой пары записей из обучающей выборки. Затем каждую такую пару можно представить как точку в пространстве признаков, в этом случае мы получим два набора точек: те, что относятся к классу соответствующих пар записей и те, что относятся к классу несоответствующих пар. Если задать расстояние для определения близости точек друг к другу, то новые пары записей можно относить к тому из двух классов, к которому они окажутся более близкими (рисунок 1).

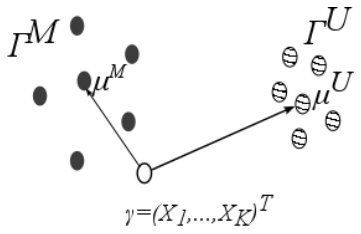

Рис. 1: Классификация на основе расстояния

Таким образом, задача выявления дубликатов сводится к задаче дискриминации двух классов: соответствующих и несоответствующих пар записей. А также классификации новых пар записей к одному из этих классов на основе функции расстояния. Данный инструмент относится к методам машинного обучения (обучение с учителем) и достаточно широко применяется при распознавании образов [17]. Более формальное описание модели представлено ниже.

4 Описание модели идентификации публикаций

Пусть даны две коллекции записей A и B. Пусть $\alpha(a)$ – запись из коллекции A, описывающий некоторый объект a; $\beta(b)$ – запись из коллекции B, описывающий объект b. В рамках данной работы объектами являются публикации, описания которых содержатся в библиографических записях.

Множество пар записей, описывающих один и тот же объект реального мира будем обозначать как M:

$$M = \langle \alpha(a), \beta(b) \rangle; a = b; \alpha(a) \in A; \beta(b) \in B. \quad (1)$$

Дополнение множества M, которое будем обозначать как U, представляет пары записей, описывающие различные объекты:

$$U = \langle \alpha(a), \beta(b) \rangle; a \neq b; \alpha(a) \in A; \beta(b) \in B. \quad (2)$$

Присвоим K признаков каждой из записей. Вектор γ содержит закодированную оценку согласованности по каждому признаку. Таким образом, γ можно представить как точку в пространстве признаков размерности K, то есть

$$\gamma = (X_1, \ldots, X_K)^T. \quad (3)$$

Для решения задачи идентификации объекта необходимо построить решающую функцию

$$D(\gamma[\alpha(a), \beta(b)]) = \begin{cases} 1, & \langle \alpha(a), \beta(b) \rangle \in M, \\ 0, & \langle \alpha(a), \beta(b) \rangle \in U, \end{cases} \quad (4)$$

служащую оценкой истинного статуса соответствия объектов

$$s(a, b) = \begin{cases} 1, & a = b, \\ 0, & a \neq b, \end{cases} \quad (5)$$

на основе имеющегося набора прецедентов.

Так называемые прецеденты – это пары $\langle \alpha(a), \beta(b) \rangle$ с известным статусом $s(a, b)$, из которых составляется обучающая выборка.

Представим обучающую выборку как два непересекающихся множества точек в пространстве признаков. Первое множество объединяет те пары записей, которые описывают один объект:

$$\Gamma^M = \{\gamma[\alpha(a), \beta(b)] | \langle \alpha(a), \beta(b) \rangle \in M\}, \quad (6)$$

второе множество включает пары, описывающие различные объекты:

$$\Gamma^U = \{\gamma[\alpha(a), \beta(b)] | \langle \alpha(a), \beta(b) \rangle \in U\}. \quad (7)$$

Тогда задача отнесения новой пары записей к одному из классов M и U может быть сведена к задаче классификации на основе вычисления некоторого расстояния до множеств Γ^M и Γ^U. Выбор расстояния обусловлен требованиями к решению задачи. В рамках данной работы в качестве расстояния предлагается использовать расстояние Махаланобиса, которое учитывает возможность взаимозависимости признаков и инвариантно к масштабу.

Квадрат расстояния Махалонобиса до центроида класса M рассчитывается согласно следующей формуле:

$$Dist^2(\gamma, \mu^M) = (\gamma - \mu^M)W^{-1}(\gamma - \mu^M)^T, \quad (8)$$

где γ - вектор значений признаков;
μ^M - центроид класса M;
W^{-1} - матрица, обратная внутригрупповой матрице ковариации.

Расстояние до центроида класса U рассчитывается аналогично:

$$Dist^2(\gamma, \mu^U) = (\gamma - \mu^U)W^{-1}(\gamma - \mu^U)^T, \quad (9)$$

где μ^U - центроид класса U.

В качестве центроида выступает вектор арифметических средних признаков, компоненты которого вычисляются по формуле:

$$\mu_i^M = \frac{1}{n^M}\sum_{k=1}^{n^M} X_{ik}^M, \quad (10)$$

где μ_i^M - i-я компонента вектора μ^M,
X_{ik}^M - значение i-й компоненты вектора $\gamma_k \in \Gamma^M$, $k = \overline{1, n^M}$.

Элементы матрицы ковариации W рассчитываются следующим образом

$$W_{ij} = \frac{1}{n^M + n^U - 2}\{\sum_{k=1}^{n^M}(X_{ik}^M - \mu_i^M)(X_{jk}^M - \mu_j^M) + \sum_{k=1}^{n^U}(X_{ik}^U - \mu_i^U)(X_{jk}^U - \mu_j^U)\}, \quad (11)$$

где n^M - число наблюдений в классе M;
n^U - число наблюдений в классе U;
X_{ik}^M - величина i-й компоненты вектора значений признаков для k-го наблюдения в классе M;
X_{ik}^U - величина i-й компоненты вектора значений признаков для k-го наблюдения в классе U;
μ_i^M - средняя величина i-й компоненты вектора значений признаков в классе M;
μ_i^U - средняя величина i-й компоненты вектора значений признаков в классе U.

В качестве критерия для построения решающей функции можно предложить минимизацию числа ошибок классификации пар из тестовой выборки

$$\min \sum_{i=1}^{N} I\{D(\gamma[\alpha(a), \beta(b)]) \neq s(a, b)\}, \quad (12)$$

где I – индикаторная функция.

5 Применение модели для библиографических данных

Предлагаемая модель базируется на методе обучения с учителем, поэтому первое и основное требование заключается в наличии обучающей выборки. Для того, чтобы составить такую выборку можно воспользоваться ISBN или ISSN. Составив выборку из таких пар записей, в которых указаны уникальные идентификаторы, можно на ее основе рассчитать весовые коэффициенты для признаков, а затем перенести эту информацию на те записи, для которых идентификаторы отсутствуют. Благодаря тому, что в ISBN и ISSN встроен механизм проверки, можно рассчитывать, что сопоставление по ним будет достаточно точным. В то же время, на практике многие записи не содержат таких идентификаторов и поэтому основывать сравнение записей только на них вряд ли возможно.

Также необходимо определить набор признаков для сравнения и способы оценки степени соответствия записей по каждому из признаков. При выборе признаков необходимо учитывать возможные вариации заполнения полей записи, возможность наличия опечаток, использование аббревиатур и т.п. Для нечеткого сравнения значений отдельных полей можно применять различные методы, такие как расстояния редактирования, метод N-грамм, хеширование по сигнатуре и т.п. Существует множество разнообразных методов нечеткого сравнения строк, а также работ, посвященных их сравнительнительному тестированию [18, 19].

В таблице 1 представлены признаки, которые предлагается использовать для сопоставления библиографических записей на книги. Для каждого признака указан один из видов сравнения текстовых значений полей: точное совпадение или нечеткое соответствие. Поскольку предлагаемый подход позволяет работать с взаимозависимыми признаками, возможно использовать различные методы нечеткого сравнения строк одновременно. В третьем столбце таблицы 1 указаны поля библиографической записи в формате RUSMARC [20], соответствующие признакам.

В таблице 1 перечислены признаки для сопоставления библиографических записей на книги. Набор признаков для периодических изданий будет аналогичным, за некоторыми исключениями. Так соответствие (out) будет вычисляться на основе ISSN, а среди признаков связанных записей (links) будут записи на то периодическое издание, в которое входит публикация.

План проведения эксперимента по выявлению дублирующихся библиографических записей выглядит следующим образом:

1. Составление обучающей и тестовой выборок:

(a) Для каждой библиографической записи в коллекции проводим поиск кандидатов в дубликаты. При этом рассматриваем только запись с указанным ISBN. Составляем набор биграмм из его заглавия (предварительно удаляем оттуда стоп-слова и неалфавитные символы);

(b) Проводим поиск записей, у которых набор биграмм (пар последовательных символов) из заглавия пересекается с текущим на 80%;

(c) Из найденных отбрасываем записи без ISBN;

(d) Из оставшихся составляем пару с текущим;

(e) Когда такие пары составлены для всех записей коллекции, полученный набор пар случайным образом разбиваем на две части - обучающую и тестовую выборки.

2. Вычисление значений признаков для каждой пары записей из обучающей выборки:

   (a) Для каждой пары, полученной на предыдущем этапе, вычисляем значения признаков, перечисленных в таблице 1.

3. Вычисление параметров системы выявления дубликатов:

   (a) Проводим анализ значимости признаков на основе их непараметрической корреляции с результирующим признаком out;

   (b) Исключаем из работы незначимые признаки;

   (c) Вычисляем оценки матрицы ковариации и центроиды классов дубликатов и недубликатов по формулам (11) и (10).

4. Тестирование:

   (a) Для каждой пары записей из тестовой выборки вычисляем расстояния до центроидов каждого из двух классов;

   (b) Относим пару к тому классу, расстояние до которого меньше;

   (c) Когда решение принято для каждой такой пары, вычисляем количество ошибочных решений и на основе него делаем вывод о качестве работы системы выявления дубликатов.

Таблица 1: Признаки для сопоставления библиографических записей на книги

Признак	Сравнение	Библиограф. запись
out (соответствие по ISBN)	точное совпадение	010$a
title (заглавие)	нечеткое соответствие	200$a
authors (авторы)	нечеткое соответствие	700$a, 700$b, 701$a, 701$b
place (место издания)	нечеткое соответствие	210$a
year (год издания)	точное совпадение	210$d
publisher (издатель)	нечеткое соответствие	210$c
edition (сведения об издании)	точное совпадение номера	205$a
pages (количество страниц)	точное совпадение	215$a
links (коды связанных записей)	точное совпадение	423 (в одной обложке с...), 461(набор)

Для того, чтобы проводить поиск по биграммам, необходимо провести индексацию и организовать точку доступа к базе данных, позволяющую такой поиск. Составление обучающей и тестовой выборок можно основывать и на отборе пар документов вручную, если есть возможность привлечения экспертов. В данной работе такая возможность отсутствует, поэтому предлагается механизм автоматического составления выборки на основе ISBN/ISSN.

6 Заключение

Предлагаемый подход обладает достаточной гибкостью в выборе признаков для сравнения записей. Благодаря тому, что в данной модели не накладывается ограничение независимости признаков, можно привлекать разнообразную информацию и даже учитывать одну и ту же информацию несколькими способами.

Ограничение предлагаемого подхода, которое заключается в необходимости построения обучающей выборки, можно обойти в том случае, если для части записей известен «ключ» по которому можно установить однозначное соответствие. В качестве такого ключа могут выступать ISBN, ISSN или другие идентификаторы источников. Возможность обучения, заложенная в предлагаемом подходе, позволяет настроиться на особенности конкретной базы библиографических записей (или нескольких баз). Это немаловажно, поскольку часто существуют такие особенности данных, обусловленные привычками каталогизаторов и внутренними правилами библиотеки.

Представленная модель была апробирована при решении задачи автоматического связывания записей разных типов: авторитетных записей, описывающих авторов, и библиографических записей, описывающих публикации [15]. Проведенные эксперименты позволили сделать вывод о применимости подхода для сопоставления записей различного типа. Это позволяет предположить, что он может использоваться и в случае с записями одного типа, то есть для выявления дубликатов.

Список литературы

[1] Рубцов Д.Н., Барахнин В.Б. Выявление дубликатов в разнородных библиографических источниках // Вестник НГУ. Сер.: Информационные технологии. – 2009. – Т. 7. – Вып.: 3. – С. 86-93.

[2] Winkler, W. E. Overview of Record Linkage and Current Research Directions. Research Report Series, RRS: Statistics #2006-2. http://www.census.gov/srd/papers/pdf/rrs2006-02.pdf.

[3] Volz J. Silk – a link discovery framework for the web of data [Electronic resource] / J. Volz [et al.] // Proc. WWW 2009 workshop on linked data on the web (LDOW 2009), Madrid, Spain, Apr. 20, 2009. – [Madrid], 2009. – 6 p. – (CEUR workshop proc. ; vol. 538). – URL: http://events.linkeddata.org/ldow2009/papers/ldow2009_paper13.pdf, free. – Tit. from the screen (usage date: 04.06.2013).

[4] Talburt J. Entity resolution and information quality / John R. Talburt. – San Francisco : Morgan Kaufmann/Elsevier, 2011. – 256 p.

[5] Jaro M. A. Probabilistic linkage of large public health data files /M. A. Jaro // Statistics in medicine. – 1995. – Vol. 14. – P. 491–498.

[6] Christen P. Febrl – freely extensible biomedical record linkage [Electronic resource] : release 0.3.1, July 1, 2005 / P. Christen, T. Churches // Austral. nat. univ. (ANU), Research school of computer sci. : [site]. – Canberra : ANU, 2013. – URL: http://cs.anu.edu.au/ Peter.Christen/Febrl/febrl-0.3/febrldoc-0.3, free. – Tit. from the screen (usage date: 04.06.2013).

[7] Jurczyk P. FRIL: a tool for comparative record linkage [Electronic resource] / P. Jurczyk [et al.] // AMIA : Annu. symp. proc. / Amer. med. informatics assoc. (AMIA). – [Bethesda] : AMIA, 2008. – Vol. 2008. – P. 440–444. – URL: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2656092/pdf/amia-0440-s2008.pdf, free. – Tit. from the screen (usage date: 04.06.2013).

[8] Bachteler T. Merge ToolBox – MTB. Getting Started [Electronic resource] : record linkage software / T. Bachteler ; German record linkage center. – Vers. 0.74. – Duisburg : [RLC], 2012 (May, 25). – 12 [3] p. – URL: http://www.uni-due.de/ hq0215/documents/mtb_gettingstarted.pdf, free. – Tit. from the screen (usage date: 04.06.2013).

[9] Hylton J. A. Identifying and merging related bibliographic records : [Electronic resource] : thes. submitted ... for the degrees of MENG in EECS and BS CSE / Jeremy A. Hylton ; Mass. Inst. of Technology (MIT), Dept. of electrical eng. and computer sci. – [Cambridge, MA : MIT], 1996. – 99 p. – (MIT-LCS-TR-678). – URL: http://publications.csail.mit.edu/lcs/pubs/pdf/MIT-LCS-TR-678.pdf, free. – Tit. from the screen (usage date: 04.06.2013).

[10] Bilenko M. Learnable similarity functions and their application to record linkage and clustering [Electronic resource] : diss. ... for the degree of DPh / Mikhail Yuryevich Bilenko ; Univ. of Texas. – Austin, 2006. – 136 p. – The electronic version of print. publ. – Access from ProQuest Dissertations and Theses. – Title from the screen.

[11] Hammerton J. On generating large-scale ground truth datasets for the deduplication of bibliographic records / J. Hammerton, M. Granitzer, D. Harvey, M. Hristakeva, K. Jack // Proceedings of the 2nd International Conference on Web Intelligence, Mining and Semantics. – ACM, 2012. – 18 p.

[12] VIAF: The virtual international authority file [Electronic resource] : [offic. site] / OCLC: the world's libraries – Dublin, 2010–2012. – URL: http://viaf.org, free. – Tit. from the screen (usage date: 04.06.2013).

[13] Bennett R. VIAF (Virtual international authority file): linking the Deutsche Nationalbibliothek and Library of Congress name authority files / R. Bennett [et al.] // Int. cataloging and bibliographic control. – 2007. – Vol. 36, № 1. – P. 12–19.

[14] Hamby R. 10 Percent Wrong for 90% Done: A Practical Approach to Collection Deduping // Computers in Libraries. – 2012. – Т. 32. – N. 4. – С. 17-21.

[15] Князева А.А., Турчановский И.Ю., Колобов О.С. Автоматическое связывание документов //Электронные библиотеки: перспективные методы и технологии, электронные

коллекции: Труды XIV Всероссийской научной конференции RCDL'2012. Переславль-Залесский, Россия, 15-18 октября 2012 г. – г. Переславль-Залесский: изд.-во «Университет города Переславля», 2012. – С. 360-369.

[16] Князева А.А. Принципы идентификации объектов в структурированных документах // Вестник НГУ. Сер.: Информационные технологии. – 2013. – Т. 11. – N 1. – С. 58-69.

[17] Ту Д., Гонсалес Р. Принципы распознавания образов : пер. с англ. / Д. Ту, Р. Гонсалес. – М. : Мир, 1978. – 411 с.

[18] Цыганов Н.Л., Циканин М.А. Исследование методов поиска дубликатов веб-документов с учетом запроса пользователя // Интернет-математика 2007: Сб. работ участников конкурса. – Екатеринбург: Изд-во Урал. ун-та, 2007. С. 211-222.

[19] Бойцов Л.М. Классификация и экспериментальное исследование современных алгоритмов нечеткого словарного поиска // Труды 6-ой Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» – RCDL2004, Пущино, Россия, 2004. http://www.rcdl.ru/papers/2004/paper27.pdf

[20] Российский коммуникативный формат (RUSMARC) [Электронный ресурс] : [сайт] / Мин-во культуры Рос. Федерации, Рос. библ. ассоц., Нац. Служба развития системы форматов RUSMARC. – [СПб., 2012]. – URL: http://www.rusmarc.ru/index.html, свободный. – Загл. с экрана (дата обращения: 04.06.2013).

Identification of duplicates in bibliographic databases

Anna A. Knyazeva, Igor Y. Turchanovsky, Oleg S. Kolobov

The problem of identification of duplicate documents in the electronic catalog of the library is considered. A model to identify duplicates based on supervised learning is proposed. The training set to configure the specific features of the database is built on the basis of those documents, which is known identifier ISBN or ISSN. Next, calculated on the basis of training sample weights are used to work with documents that do not have ISBN or ISSN.