УДК 004.622’417

А. А. Князева

Томский филиал Института вычислительных технологий СО РАН
пр. Академический, 10/4, Томск, 634055

E-mail: aknjazeva@ict.nsc.ru

ПРИНЦИПЫ ИДЕНТИФИКАЦИИ ОБЪЕКТОВ В СТРУКТУРИРОВАННЫХ ДОКУМЕНТАХ

Рассматривается задача идентификации объектов реального мира, упоминаемых в структурированных документах. Сформулированный подход позволяет учитывать различные признаки, по которым производится идентификация, и присваивать им различные веса в зависимости от их значимости. Рассматривается применение предлагаемой модели к задаче идентификации персон, выступающих в роли авторов публикаций, на основе данных электронного каталога библиотеки.

Ключевые слова: идентификация объектов, базы данных, структурированные документы, связывание записей.

Введение

Проблема идентификации объектов в различных информационных ресурсах в последнее время проявляется все более остро. Такая идентификация помогает существенно повысить качество информационного поиска за счет повышения его полноты и точности [1]. Идентификация персон дает возможность избежать несоответствий при смене фамилии, места работы и т. д. Идентификация организаций позволяет учитывать их переименования и переформирования. Суть задачи идентификации заключается в том, чтобы определить, в каких документах идет речь об одном и том же объекте реального мира, и установить связь между этими документами [2; 3]. Целью данной работы является разработка единого подхода к идентификации объектов и формулировка соответствующей модели процесса идентификации.

В качестве документа может выступать элемент информации, хранящейся в базе данных, например, кортеж в реляционной базе данных или полнотекстовый документ с внедренными метаданными. Главное требование к документу – он должен содержать информацию о свойствах объекта в виде набора атрибутов с определенной структурой. Термин «документ» выбран для того, чтобы подчеркнуть применимость приведенных методов не только к записям базы данных, но и к любым структурированным документам, содержащим информацию о некоторых объектах. В качестве примера можно привести применение приведенной ниже модели к полнотекстовым документам, хранящимся в электронных библиотеках, при условии предварительной обработки для выявления значений необходимых атрибутов (имя автора, заглавие и т. п.). При необходимости и сам текст такого документа может рассматриваться в качестве атрибута. Таким образом, под документом в таком контексте понимается совокупность сведений об одном или нескольких объектах, представленная как структурированный набор данных. В качестве объектов могут выступать персоны, организации, географические места и др.

Князева А. А. Принципы идентификации объектов в структурированных документах // Вестн. Новосиб. гос. ун-та. Серия: Информационные технологии. 2013. Т. 11, вып. 1. С. 58–67.

Существуют и другие термины для названия данной задачи. Так, в приложениях для работы с базами данных ее принято называть «процедурой слияния / чистки», «очищением списка» и т. п., в области информационных технологий часто встречаются такие названия, как «сопоставление данных», «идентификация экземпляров», «разрешение сущностей», «разрешение перекрестных ссылок», «выравнивание данных» и др.

В более общей формулировке задача идентификации может быть поставлена для документов разных типов, имеющих различную структуру. Задача связывания документов одного типа, т. е. выявления дублирующихся документов в одном или нескольких источниках, представляет ее частный случай. При этом речь идет о нечетких дубликатах, поскольку нередки ситуации, когда дублирующиеся документы имеют различные значения в одном или нескольких полях [4]. Причинами такого несоответствия могут быть опечатки, транспозиции символов, измененный порядок слов, использование сокращений и аббревиатур, разночтения в зарубежных транскрипциях, неполнота данных и т. п. [5].

Безусловно, самым простым подходом к рассматриваемой задаче было бы принятие решения о соответствии документов на основе некоторых правил, которые могут быть относительно простыми или достаточно сложными в зависимости от конкретной системы. Такой подход к установлению связей можно назвать детерминистическим или эмпирическим. Однако на практике далеко не всегда есть возможность выработать исчерпывающий набор правил, особенно в условиях наличия пропусков в данных.

Впервые задача автоматического связывания без применения фиксированных правил была сформулирована Ньюкомби [6] в контексте сопоставления записей о рождениях с записями о регистрации брака. Суть предложенного решения заключается в подсчете количества совпавших полей. Если это количество превышает некоторый заданный заранее порог, то записи признаются соответствующими, в противном случае – несоответствующими. В дальнейшем для идей Ньюкомби была разработана формальная математическая модель, получившая название вероятностной модели связывания, основанной на ошибках [7], на которой в настоящее время базируется целое семейство вероятностных моделей, например, модели, основанные на штрафах или использующие EM-алгоритм [4]. Описанный подход основан на явной оценке условных вероятностей соответствия записей, он предполагает знание распределения признаков соответствия или их взаимную независимость [8].

Альтернативой является более прямой подход, основанный на методиках машинного обучения [9]. Это может быть обучение с учителем или без него. Основная идея заключается в том, чтобы относить пару документов к классу соответствующих или несоответствующих пар на основании ее схожести с остальными парами класса. В рамках данной работы используется классификация на основе расстояния Махаланобиса [10].

Модель

Пусть даны две коллекции документов A и B. Пусть $\alpha(a)$ – документ из коллекции A, описывающий некоторый объект a; $\beta(b)$ – документ из коллекции B, описывающий объект b.

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

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

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

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

Присвоим K признаков каждому из документов. Вектор γ содержит закодированную оценку согласованности по каждому признаку. Таким образом, γ можно представить как точку в пространстве признаков размерности K, т. е. $\gamma = (X_1, ..., X_K)^T$.

Для решения задачи идентификации необходимо построить решающую функцию

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

служащую оценкой истинного статуса соответствия объектов

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

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 документов, которые описывают один объект:

$$\Gamma^M=\left\{\gamma\left[\alpha(a),\beta(b)\right]\middle|\langle\alpha(a),\beta(b)\rangle\in M\right\}.$$

Второе множество включает пары, описывающие различные объекты:

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

Тогда задача отнесения новой пары документов 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:

$$\mathrm{Dist}^2\left(\gamma,\mu^M\right)=\left(\gamma-\mu^M\right)W^{-1}\left(\gamma-\mu^M\right)^T,$$

где

γ – вектор значений признаков;

μ^M – центроид класса M;

W^{-1} – матрица, обратная внутригрупповой матрице ковариации.

Расстояние до центроида класса U рассчитывается аналогично:

$$\mathrm{Dist}^2\left(\gamma,\mu^U\right)=\left(\gamma-\mu^U\right)W^{-1}\left(\gamma-\mu^U\right)^T,$$

где μ^U – центроид класса U.

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

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

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:

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

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

$$\min \sum I\left\{D\left(\gamma\left[\alpha(a),\beta(b)\right]\right) \neq s(a,b)\right\},$$

где I – индикаторная функция.

Пример применения модели

В качестве применения предложенной модели была рассмотрена задача идентификации персон, упоминаемых в электронном каталоге библиотеки [11; 12]. В качестве коллекции B выступает база библиографических документов, содержащая описания публикаций, а в качестве коллекции A – база авторитетных документов имен авторов [13].

В такой постановке задачи существуют некоторые особенности. В документах коллекции B может упоминаться сразу несколько персон, если они являются соавторами публикации, тогда как каждый документ из коллекции A посвящен описанию одной персоны. Таким образом, к описанной выше задаче добавляется такое ограничение: для $\forall$ объекта a $\exists$ не более 1 документа $\alpha(a)$, $\alpha(a) \in A$ и может существовать несколько документов $\beta(a)$, $\beta(a) \in B$. Таким образом, для идентификации персоны a, упоминаемой в документе $\beta(a)$, необходимо и достаточно связать этот документ с одним и только одним документом $\alpha(a)$. Документ $\alpha(a)$ будем называть авторитетным или нормативным, поскольку он однозначно указывает на объект.

Для того чтобы реализовать описанную модель в виде алгоритма идентификации персон на этапе загрузки документа β в базу данных B, можно разделить процесс идентификации объектов на этапы, за каждый из которых будет отвечать соответствующий функциональных блок:

1) подготовка данных;
2) составление пар;
3) сравнение отдельных полей в парах документов;
4) решающая функция.

Кроме этих четырех этапов, непосредственно участвующих в процедуре связывания, необходимо наличие еще двух: настройка системы и проверка качества идентификации. Последние два включаются в работу периодически при расширении базы данных. Принцип работы у них общий: для документа, относительно которого уже известно правильное решение (с каким из авторитетных документов он должен быть связан), проводится процедура идентификации, и в первом случае уточняются параметры системы, а во втором оценивается, насколько успешно система справилась с задачей.

Документ β, загружаемый в базу данных в процессе идентификации, может находиться в одном из четырех возможных состояний.

- Документ в том виде, в котором он поступает на вход процедуры, отметка о связи отсутствует – $\beta^{(0)}$.
- Документ прошел предварительную подготовку и корректировку отдельных полей – $\beta^{(1)}$.
- Документ находится на дополнительном рассмотреннии, поскольку для него было подобрано более одного подходящего документа $\alpha - \beta^{(2)}$.
- Документ содержит явное указание на соответствующий документ $\alpha - \beta^{(3)}$.

Переходы между этими состояниями отображены на рис. 1.

Рассмотрим подробнее описанные выше этапы.

Блок подготовки (рис. 2) позволяет при необходимости очистить входной документ β от ошибок, недопустимых значений и т. п.

Кроме того, на этапе подготовки осуществляется проверка на предмет наличия достаточного количества информации для идентификации. Для проведения такой проверки необходимо сформулировать входные требования, задающие минимальный набор полей, достаточный для работы [14].

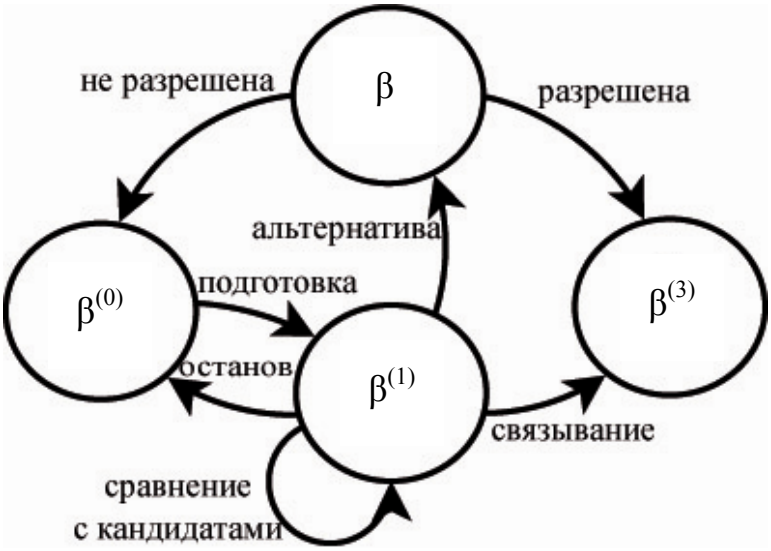

Рис. 1. Состояния документа β

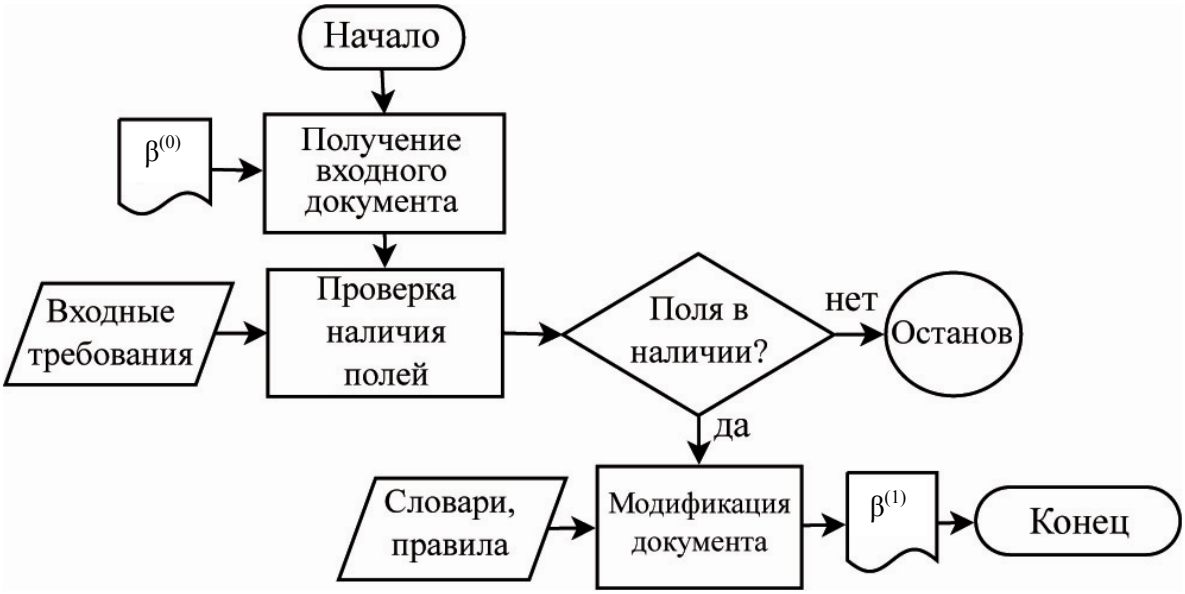

Рис. 2. Модель блока подготовки

Блок составления пар. Сравнение входящего документа β с каждым из авторитетных документов α может оказаться достаточно трудоемким процессом, особенно если осуществлять его «на лету». Необходим механизм сокращения количества авторитетных документов, которые будут сопоставляться с входящим. Такой механизм можно реализовать в виде отдельного функционального блока, отвечающего за составление пар документов (рис. 3).

В рамках данной работы был принят метод поиска по составному ключу, состоящему из двух значений: фамилия и инициалы автора. Значение ключа определяется по входящему документу β, а поиск производится в коллекции *A*. При этом используется точное сопостав-

ление. Такой механизм позволяет существенно снизить трудоемкость без использования сложных вычислений.

Одной из важных черт предлагаемого подхода является использование расширенного авторитетного документа (рис. 4) для сравнения с входящим документом. Аналогичный подход используется в проекте VIAF [15]. Расширенный авторитетный документ, кроме самого найденного авторитетного документа, включает информацию из библиографических документов, уже хранящихся в системе и связанных с ней.

Такой подход позволяет увеличивать объем информации, задействованной в анализе, и получать более точные результаты.

Блок сравнения отдельных полей в паре документов. Цель блока сравнения отдельных полей (рис. 5) заключается в оценке того, насколько документы совпадают по различным параметрам. Результатом работы блока является вектор, составленный из оценок близости двух строк, которые являются значениями соответствующих полей.

В рамках настоящей работы используется комбинация точного сравнения и сравнения с усечением, определяемого с помощью стеммера Портера для русского языка [1]. Сравнение полей производится для каждой из пар документов, полученных в результате работы блока составления пар.

Блок принятия решения. Соответствие на уровне записей необязательно означает однозначное соответствие на уровне полей. Для принятия решения о соответствии в рамках данной работы используется индукционная модель. Классификация пары документов к классу соответствующих либо несоответствующих пар производится с помощью расстояния Махалонибиса до каждого из двух классов, определенных с помощью обучающей выборки.

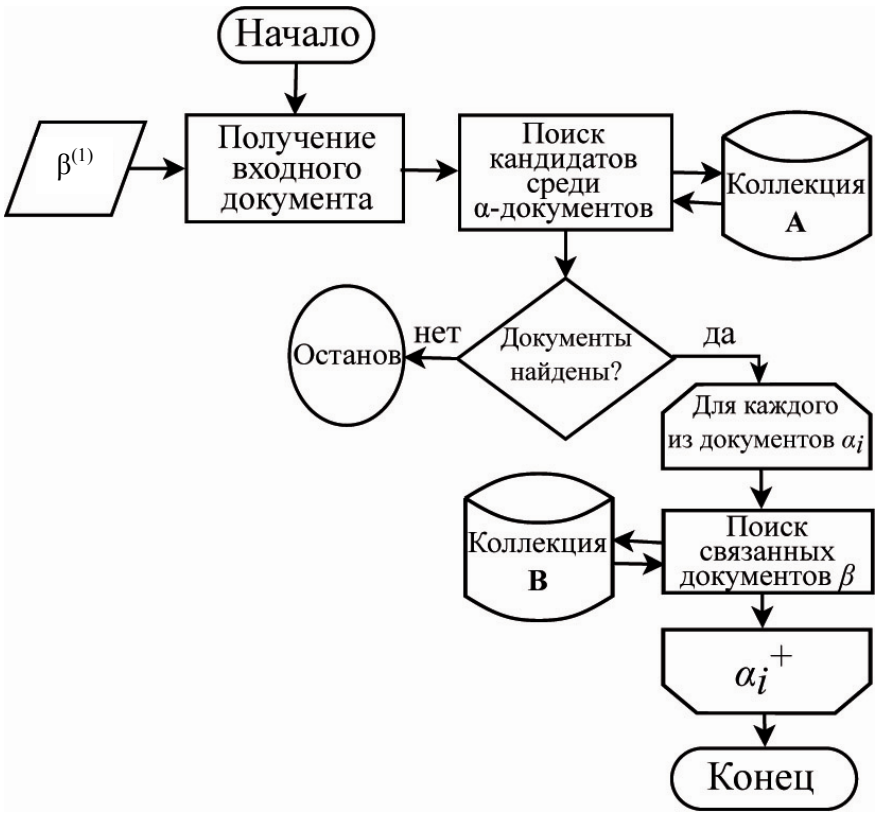

Рис. 3. Модель блока составления пар

[1] Russian stemming algorithm. URL: http://snowball.tartarus.org/algorithms/ russian/stemmer.html

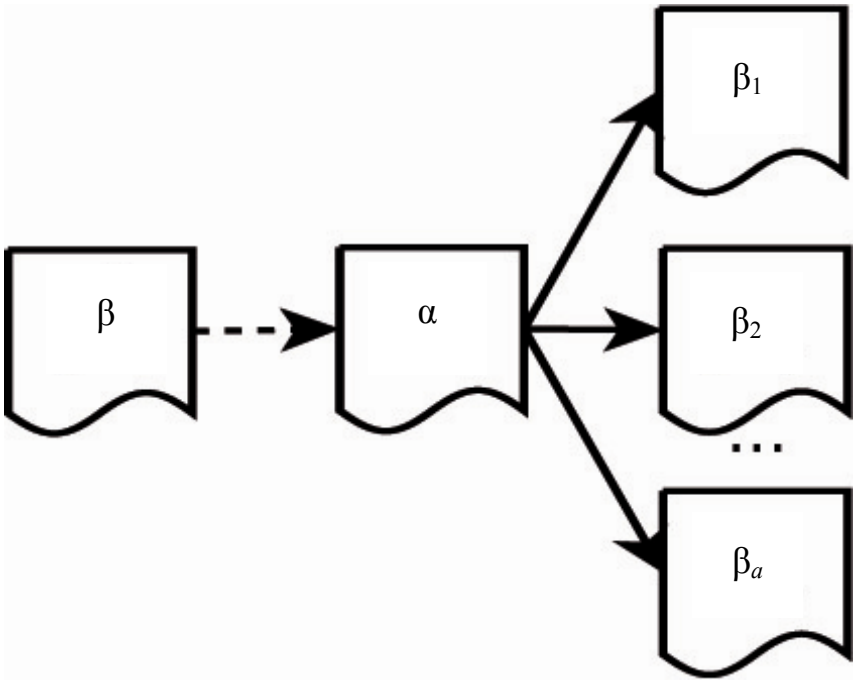

Рис. 4. Расширенный авторитетный документ

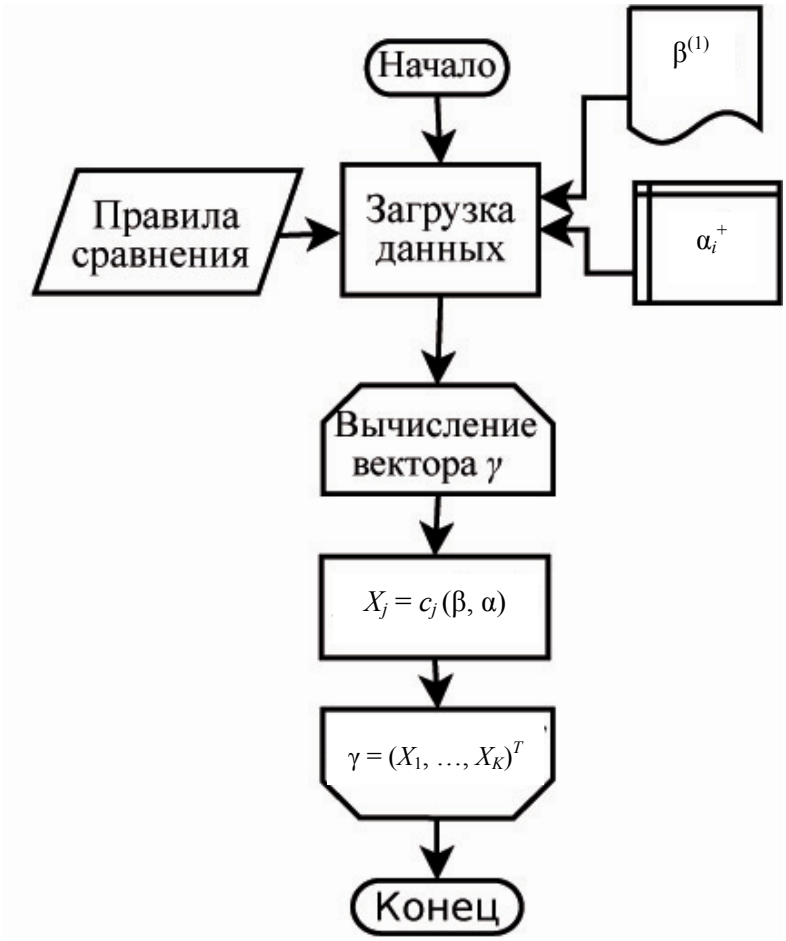

Рис. 5. Модель блока сравнения полей

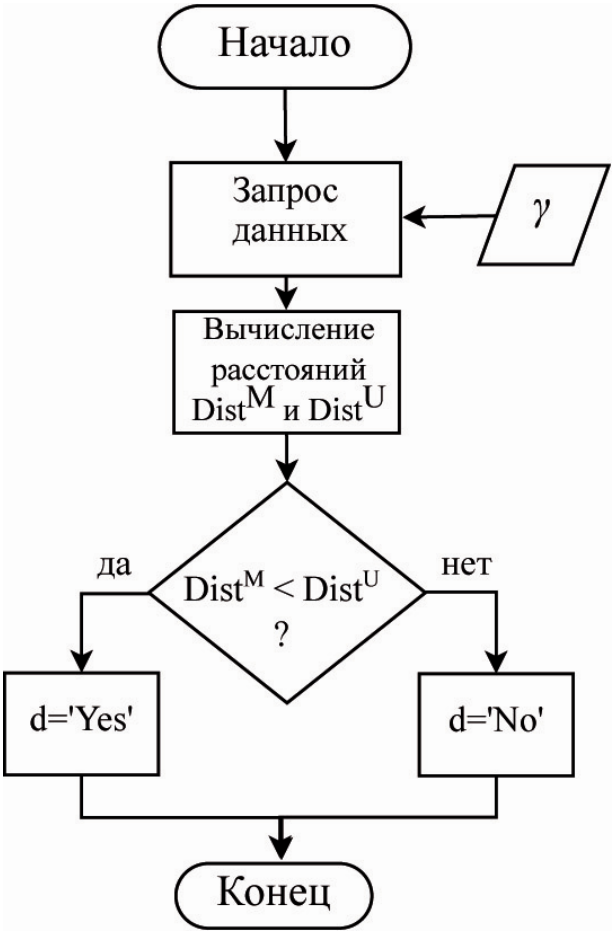

Рис. 6. Процедура принятия частного решения

Само решение о выборе документа α, с которым следует связать входной документ β, можно разделить на две части.

1. Частное решение о соответствии, которое принимается индивидуально для каждой из составленных пар документов (рис. 6).

2. Общее решение о том, был ли среди кандидатов подходящий документ α, с которым следует установить связь.

Общее решение призвано избежать такой ситуации, когда документ β связывается сразу с несколькими документами α, поскольку она противоречит постановке задачи.

Принятие окончательного решения. Итак, когда для всех пар ⟨α, β⟩ приняты частные решения о соответствии, требуется определить, как именно следует поступить с документом: внести в него отметку о соответствующем α или оставить его без изменений. Пусть I_D – индикаторная функция, определенная следующим образом:

$$I_D(D_i) = \begin{cases} 1, D_i = \text{'Yes'}, \\ 0, D_i = \text{'No'}. \end{cases}$$

Тогда значение суммы $\sum_{i=1}^{k} I_D(D_i)$ равно количеству положительных решений в наборе частных решений (D_1, D_2, ..., D_k). На основе этого количества и принимается окончательное решение:

$$\tilde{D} = \begin{cases} \tilde{D}_-, \sum_{i=1}^{k} I_D(D_i) = 0, \\ \tilde{D}_+, \sum_{i=1}^{k} I_D(D_i) = 1, \\ \tilde{D}_0, \sum_{i=1}^{k} I_D(D_i) > 1. \end{cases}$$

В приведенном выражении $\tilde{D}_-$ означает, что для рассматриваемого β не было найдено ни одного α, соответствующего ему. В этом случае β остается в состоянии $\beta^{(0)}$ и может снова поступить на вход процедуры связывания позднее.

Решение $\tilde{D}_+$ означает, что β был поставлен в соответствие α, при этом в β вносится отметка об установленной связи и он переходит в состояние $\beta^{(3)}$.

Последний вариант – решение $\tilde{D}_0$ – отвечает ситуации, когда для β было найдено более одного α и для разрешения возникшей коллизии необходимо участие эксперта. В этом случае отметок о связи в β не делается, и он переходит в состояние $\beta^{(2)}$ для дальнейшего анализа.

Заключение

В работе представлена модель идентификации объектов реального мира, упоминаемых в структурированных документах, на основе сопоставления указанных в них признаков объектов. Рассматриваемая модель не требует принятия предположений о распределении признаков и позволяет учесть их взаимозависимость. Также в модели учитываются уже установленные связи между документами коллекций. Область применения предложенной модели достаточно широка. В качестве частного случая модель может использоваться для решения задачи выявления дублирующихся документов в коллекции.

В качестве применения предложенной модели рассматривается задача идентификации персон, указанных как авторы публикаций в электронном каталоге. Такая идентификация позволяет избежать путаницы между публикациями однофамильцев, учесть смену фамилии, псевдонимы и разночтения в транскрипциях иностранных фамилий.

Ограничение представленной модели заключается в необходимости обучающей выборки, состоящей из пар документов, для которых известно соответствие. Такая выборка позволяет настроиться на особенности конкретной коллекции и придать больший вес тем признакам, которые более значимы. Использование обучающей выборки позволяет отказаться от разработки эмпирических правил для связывания документов.

Список литературы

1. *Князева А. А., Колобов О. С., Турчановский И. Ю., Федотов А. М.* Ранжированный поиск в библиографических базах данных // Вестн. Новосиб. гос. ун-та. Серия: Информационные технологии. 2009. Т. 7, вып. 4. С. 81–96.

2. *Князева А. А., Турчановский И. Ю., Колобов О. С.* Автоматическое связывание документов // Электронные библиотеки: перспективные методы и технологии, электронные коллекции: Тр. XIV Всерос. науч. конф. RCDL’2012. Переславль-Залесский: Изд-во «Университет города Переславля», 2012. С. 360–369.

3. *Князева А. А., Турчановский И. Ю., Колобов О. С.* Автоматическое связывание структурированных документов // Материаловедение, технологии и экология в 3-м тысячелетии: Сб. докл. V Всерос. конф. молодых ученых [Электронный ресурс]. Томск: Изд-во ИОА СО РАН, 2012. CD-ROM.

4. *Elfeky M. G., Elmagarmid A. K., Verykios V. S.* TAILOR: A Record Linkage Tool Box // Proc. of the XVIII International Conference on Data Engineering (ICDE 02). IEEE Computer Societyю Washington, DC, 2002. P. 17–28.

5. *Рубцов Д. Н., Барахнин В. Б.* Выявление дубликатов в разнородных библиографических источниках // Вестн. Новосиб. гос. ун-та. Серия: Информационные технологии. 2009. Т. 7, вып. 3. С. 86–93.

6. *Newcombe H. B., Kennedy J. M., Axford S. J., James A. P.* Automatic Linkage of Vital Records // Science. 1959. Vol. 130. P. 954–959.

7. *Fellegi I. P., Sunter A. B.* A Theory for Record Linkage // J. of the American Statistical Association. 1969. Vol. 64. P. 1183–1210.

8. *Belin T. R., Rubin D. B.* A Method for Calibrating False-Match Rates in Record Linkage // J. of the American Statistical Association. 1995. Vol. 90. P. 694–707.

9. *Bilenko M., Mooney R.* Learning to Combine Trained Distance Metrics for Duplicate Detection in Databases: Technical Report AI-02-296 / Artificial Intelligence Lab. University of Texas at Austin, 2002.

10. *Mahalanobis P. C.* On the Generalized Distance in Statistics // Proc. of the National Institute of Sciences of India. 1936. Vol. 2 (1). P. 49–55.

11. *Князева А. А., Турчановский И. Ю., Колобов О. С.* Автоматический авторитетный контроль для распределенных библиографических баз данных // Распределенные информационные и вычислительные ресурсы (DICR'2010): Материалы XIII Рос. конф. с участием иностранных ученых [Электронный ресурс]. Новосибирск: ИВТ СО РАН, 2010. CD-ROM.

12. *Князева А. А., Колобов О. С.* Восстановление связей между библиографическими записями // Современные проблемы математики, информатики и биоинформатики: Материалы Междунар. конф., посвящ. 100-летию со дня рождения члена-корреспондента АН СССР Алексея Андреевича Ляпунова [Электронный ресурс]. Новосибирск: ИВТ СО РАН, 2011. CD-ROM.

13. *Федотов А. М., Жижимов О. Л., Князева А. А., Колобов О. С., Мазов Н. А., Турчановский И. Ю., Федотова О. А.* Проблемы авторитетного контроля для распределенных электронных библиотек и библиографических баз // Вестн. Новосиб. гос. ун-та. Серия: Информационные технологии. 2011. Т. 9, вып. 1. С. 89–101.

14. *Князева А. А., Колобов О. С., Турчановский И. Ю.* Наличие информации для связывания на примере базы данных «MedArt» // Распределенные информационные и вычислительные ресурсы (DICR'2012): Материалы XIV Рос. конф. с междунар. участием [Электронный ресурс]. Новосибирск: ИВТ СО РАН, 2012. CD-ROM.

15. *Bennett R., Christal H.-D., O'Neill E. T., Tillett B.* VIAF (Virtual International Authority File): Linking the Deutsche Nationalbibliothek and Library of Congress Name Authority Files // International Cataloging and Bibliographic Control. 2007. Vol. 36 (1). P. 12–19.

Материал поступил в редколлегию 12.02.2013

A. A. Knyazeva

PRINCIPLES OF IDENTIFICATION OF OBJECTS IN STRUCTURED DOCUMENTS

The paper describes the problem of real word objects identification, which are mentioned in the structured documents. The approach takes into account different features for identification and its weights depending on its significance. The application of the proposed model to the problem of identification of persons that act as authors of publications based on data from the electronic library catalog is considered.

Keywords: identification of objects, databases, structured documents, record linkage.