Модели документального и фактографического поиска в электронных библиотеках

© В.Б. Барахнин © А.М. Федотов
Институт вычислительных технологий СО РАН,
Новосибирский государственный университет
bar@ict.nsc.ru fedotov@sbras.ru

Аннотация

В статье рассматриваются вопросы построения моделей документального и фактографического поиска в электронных библиотеках, работающих с документами достаточно произвольной структуры. Предложена модель классификации документов электронной библиотеки, основанная на использовании отношения толерантности, учитывающая возможное отсутствие априорно заданных классификаторов. Показано, что при создании фактографических систем целесообразно следующее понимание факта: содержащаяся в тексте и метаданных документа совокупность связей между сущностями, описываемыми в онтологии информационной системы. Предложена простейшая модель онтологии фактографической системы.

Работа выполнена при частичной поддержке РФФИ (проекты 12-07-00472, 13-07-00258), президентской программы «Ведущие научные школы РФ» (грант НШ 5006.2014.9) и интеграционных проектов СО РАН.

1 Введение

В работе [1] была предложена модель электронной библиотеки (ЭБ) по научному наследию, выступающей в качестве основы интеллектуальной системы (ИнтС), предназначенной не только для документального, но и для фактографического поиска, то есть позволяющей удовлетворять информационные потребности квалифицированного пользователя в соответствии со схемой «документ – факт – рассуждение» [2, 3]. При этом, как отмечено в [3], важной проблемой является построение моделей основных компонентов интеллектуальной системы: как информационно-поисковой системы (рассматриваемой в абстрактном виде, то есть без учета средств технической реализации), так и логических компонент, отвечающих за поиск информации, вывод новых знаний и диалог с пользователем.

В настоящее время большинство работ, направленных на решение указанной задачи, исходит из неявного предположение о возможности широкого распространения более или менее подробной стандартизации представления информации, например на основе словарей, как это сделано в рамках концепции Semantic Web консорциума W3 [4].

Однако попытки автоматизировать процессы обработки реальных массивов документов, например, размещенных в сети Интернет, и извлечения из них фактографической информации использование концепции Semantic Web неизбежно порождает серьезные проблемы, поскольку наработки консорциума W3 носят лишь рекомендательный характер, а объявить их стандартами могут только организации, имеющие соответствующий статус, например ISO, ГОСТ или ANSI. Ввиду этого реальное развитие большинства ресурсов Интернета, в том числе научной направленности, идет без учета подобных необязательных рекомендаций. Более того, свободный характер размещения материалов в сети Интернет превращает требование соблюдения даже обязательных стандартов представления информации всего лишь в благое пожелание (особенно это касается российской части Интернета). Разумеется, сказанное относится еще в большей степени к электронным документам, не размещенным в Интернете и полученным создателями ИнтС для обработки посредством локального доступа.

Таким образом, возникает необходимость разработки моделей документального и фактографического поиска в электронных библиотеках, работающих с документами достаточно произвольной структуры.

Труды 16-й Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» — RCDL-2014, Дубна, Россия, 13–16 октября 2014 г.

2 Модель классификации документов

Так как задачи поиска и классификации информации взаимно-обратны, то нам достаточно рассмотреть модель классификации документов, наиболее адекватно отражающую особенности работы с ЭБ, в частности, возможное отсутствие априорно заданных классификаторов.

Наиболее распространенным вариантом классификации библиографических ресурсов является фасетная классификация, теория построения которой формализована индийским библиотековедом Ш.Р. Ранганатаном (см. [5]). Объекты классифицируются одновременно по нескольким независимым друг от друга признакам (фасетам). Применительно к ЭБ (и электронным ресурсам вообще) в качестве фасетов выступают элементы метаданных.

Важно отметить, что при создании научно-образовательных ЭБ, для которых библиографические признаки документов гораздо менее важны по сравнению с обычными ЭБ, подмножества множеств значений библиографических метаданных, образующих значения фасетов, как правило, более широки. Так, ссылки на различные переиздания одного и того же документа с точки зрения научно-образовательных ЭБ целесообразно считать эквивалентными.

Простейшая формальная модель классификации документов с использованием структурированных метаданных документов выглядит следующим образом [6]. Пусть в справочно-поисковом аппарате ЭБ хранится информация о документах d_i. При этом любой документ d_i представляется как $d_i = <m_i^{j,k}>$, где $m_i^{j,k}$ – значения элементов метаданных M^j, k – количество значений (с учетом повторений) соответствующего элемента метаданных в описании документа. Рассмотрим подмножество метаданных M_C, определяющее набор классификационных признаков документов, используемых для составления поискового предписания (с учетом заданных логических операций). Для фиксированного элемента метаданных M^j, где $M^j \subset M_C$, заранее определяются подмножества M_i^j множества значений этого элемента метаданных (указанные подмножества могут, вообще говоря, пересекаться).

Будем считать два документа *толерантными* (напомним, что толерантность – отношение, которое обладает свойствами рефлексивности и симметричности, но, вообще говоря, может не обладать, в отличие от отношения эквивалентности, свойством транзитивности; подробно свойства этого отношения исследованы в [7]), если у них значения некоторого элемента метаданных входят в одно и то же подмножество M_i^j, при этом если значения рассматриваемого элемента метаданных могут повторяться, то документы считаются толерантными при совпадении хотя бы одного из значений. Каждое такое подмножество порождает на множестве документов ЭБ предкласс толерантности, который обозначим K_i^j.

Более того, в большинстве случаев такие предклассы максимальны, т.е. являются классами толерантности. Предкласс K_k^i является классом, если не существует отличного от него (т.е. порожденного другим набором элементов метаданных) предкласса K_l^j такого, что $K_k^i \subset K_l^j$, в противном случае K_k^i классом не является.

Выясним, в каких случаях предклассы не являются классами (это необходимо, например, для описываемого ниже определения базиса пространства толерантности). Прежде всего, если $M_l^i \subset M_k^i$, то $K_k^i \subseteq K_l^i$, и поэтому K_k^i классом не является, за исключением конкретного подбора документов, когда $K_k^i = K_l^i$, но и в этом случае, очевидно, нет смысла рассматривать K_k^i в качестве отдельного класса. С содержательной точки зрения этой ситуации соответствует вхождение некоторого раздела классификатора ЭБ в раздел более высокого уровня, когда оба этих раздела учитываются при описании пространства толерантности (разумеется, можно и не учитывать раздел более низкого уровня при определении толерантных элементов, но тогда мы будем иметь дело с пространством толерантности, отличным от первоначального). В описанной ситуации предклассы, не являющиеся классами, определяются априори.

Однако возможна и ситуация, когда $K_k^i \subset K_l^j$ из-за конкретных особенностей документов ЭБ. Например, в ЭБ по истории математики все документы, имеющие географический признак *Египет*, имеют хронологической признак *до новой эры*, при этом указанный хронологический признаки имеют и документы, относящиеся к другим регионам. Ясно, что в этом случае все документы с признаком *Египет* попарно толерантны не только в силу географического, но и в силу хронологического признака, однако появление в ЭБ хотя бы одного документа с признаком *Египет*, датируемого *новой эрой*, изменит эту ситуацию. Тем самым в рассматриваемой ситуации предкласс K_k^i целесообразно рассматривать (например, при построении базиса) в качестве класса.

Совокупность всех классов толерантности (включая предклассы, рассматриваемые в соответствии со сказанным выше в качестве классов) будем обозначать через H.

Далее опишем, как устроен базис описываемого пространства толерантности (некоторая совокупность H_B классов толерантности называется базисом, если для всякой толерантной пары документов существует класс из H_B, содержащий оба этих документа, а удаление из H_B хотя бы одного класса приводит к потере этого свойства). Очевидно, что множество классов толерантности H_M (включающее по нашему

построению, в том числе, и предклассы, рассматриваемые в качестве классов), порожденных всей совокупностью подмножеств M_i^j, содержит базис. Утверждать, что H_M в точности является базисом нельзя потому, что входящие в него предклассы, не являющиеся классами, могут быть удалены без потери первого свойства из определения базиса. Однако, поскольку добавление в ЭБ даже одного документа может сделать предкласс классом и, стало быть, «полноценным» элементом базиса, постольку рассмотрение таких предклассов в качестве элементов базиса целесообразно с точки зрения организации классификации и поиска документов ЭБ.

Описание классов толерантности для ЭБ имеет большое практическое значение. Прежде всего, рассмотрим множество всех документов, для которых существует такая совокупность классов (включая предклассы, рассматриваемые в качестве классов) из H, что каждый из этих документов входит в эти и только эти классы. Такое множество представляет собой ядро толерантности, а множество всех ядер толерантности задает отношение эквивалентности на множестве документов ЭБ. При этом для построения ядер толерантности достаточно рассматривать лишь классы (и предклассы) из базиса H_M [7].

Таким образом, поисковое предписание, содержащее подмножество метаданных, определяющее набор классификационных признаков, с указанием сочетаний значений этих метаданных при помощи логических операций, определяет конкретное ядро толерантности на множестве документов, которое и выдается пользователю в качестве ответа на его информационный запрос.

Кроме того, на множестве классов толерантности также можно, в свою очередь, ввести отношение толерантности, при этом толерантными считаются классы, имеющие хотя бы один общий документ. Такая конструкция оказывается полезной, например, для организации поиска документов «по аналогии».

Формализм, основанный на использовании отношения толерантности, оказывается более удобным при создании ЭБ, поскольку в отличие от обычных библиотек, в которых классификаторы заданы априорно, при работе с ЭБ нередко приходится использовать те или иные алгоритмы кластеризации документов (см., например, [3]), а уже потом, исходя из результатов кластеризации, устанавливать подмножества множеств значений элементов метаданных, выступающих в качестве значений фасетов.

3 Уточнение понятия «факт»

Прежде чем обсуждать проблемы работы с фактографической информацией, следует уточнить, какое именно содержание мы будет вкладывать в понятие «факт».

К сожалению, в официальных документах: ГОСТ 7.73–96 «Поиск и распространение информации» и ГОСТ 7.74–96 «Информационно-поисковые языки» – этот термин практически не формализован. Так, в ГОСТ 7.74–96 дано лишь косвенное, причем не слишком содержательное, определение факта: «7.7. **фактографическое индексирование:** Индексирование, предусматривающее отражение в поисковом образе документа конкретных сведений (фактов)». Интересно отметить, что иноязычные эквиваленты терминов, относящихся к фактографическому поиску (в отличие от подавляющего большинства прочих терминов), в указанном ГОСТе отсутствуют. Что же касается ГОСТ 7.73–96, то интересующее нас понятие косвенно раскрывается в следующем определении: «3.3.7. **база первичных данных; фактографическая база данных:** База данных, содержащая информацию, относящуюся непосредственно к предметной области». В зарубежной литературе, посвященной разработке фактографических систем, при определении понятия «элементарный факт» наиболее распространен подход, восходящий к работе [8] и оперирующий в терминах модели «объект – роль», которая, по мнению ее автора, обладает более естественными и выразительными средствами по сравнению с моделью «сущность – связь». Фактически же в этой работе, как, например, и в более современной работе [9], обсуждение указанного понятия переводится на довольно абстрактный уровень концептуального моделирования. Это значительно снижает объявленную в [8] естественность и выразительность определения, а, главное, оставляет без внимания такие важные вопросы, как непосредственное отношение тех или иных фактов (данных) к предметной области (что является важной частью процитированного выше определения фактографической базы данных из ГОСТ 7.73–96) и достоверность фактов (в классической монографии [10] достоверность названа одним из определяющих признаков факта, и, хотя в области гуманитарных наук некое утверждение, достоверность которого неоднозначна, снабженное ссылкой на источник информации, нередко становится новым утверждением, являющимся предметом изучения источниковедения, т.е. отдельным самостоятельным фактом, но полностью абстрагироваться от проблемы достоверности факта вряд ли возможно).

Для уточнения смысла, вкладываемого в термин «факт» применительно к той области информатики, которая изучает процессы взаимных преобразований данных, информации и знаний в процессе функционирования ИнтС, представляется целесообразным использование семиотического подхода. Понятие «факт» является центральным в «Логико-философском трактате» Л. Витгенштейна [11], одним из источников которого, как отметил Витгенштейн предисловии трактата, стали работы

Г. Фреге – основателя семиотики. Процитируем основные положения трактата, касающиеся фактов:

«…1.1. Мир есть совокупность фактов, а не вещей. ...

1.2. Мир распадается на факты.

1.21. Любой факт может иметь место или не иметь места, а все остальное останется тем же самым. ...

2. То, что имеет место, что является фактом, – это существование атомарных фактов.

2.01. Атомарный факт есть соединение объектов (вещей, предметов).

2.011. Для предмета существенно то, что он может быть составной частью атомарного факта. …

2.034. Структура факта состоит из структур атомарных фактов.

2.04. Совокупность всех существующих атомарных фактов есть мир.

2.05. Совокупность всех существующих атомарных фактов определяет также, какие атомарные факты не существуют.

2.06. Существование или несуществование атомарных фактов есть действительность. (Существование атомарных фактов мы также называем положительным фактом, несуществование – отрицательным.)

2.061. Атомарные факты независимы друг от друга.

2.062. Из существования или несуществования какого-либо одного атомарного факта нельзя заключать о существовании или несуществовании другого атомарного факта. …

4.21. Простейшее предложение, элементарное предложение, утверждает существование атомарного факта. …

4.22. Элементарное предложение состоит из имен. Оно есть связь, сцепление имен».

Положения, выдвинутые в «Логико-философском трактате», имеют большое значение для семиотики, в частности, потому, что в нем устанавливается полное соответствие между онтологическими и семантическими понятиями [12]. Кроме того, Витгенштейн не исключает ложные (или, если угодно, представляющиеся на данном уровне познания ложными) утверждения из числа атомарных фактов, а называет такие факты несуществующими.

Нетрудно заметить, что процитированные положения «Логико-философском трактата» (прежде всего, ключевые определения из раздела 2.01: «**Атомарный факт есть соединение объектов (вещей, предметов)... Структура факта состоит из структур атомарных фактов**») практически полностью воспроизводятся в модели данных «сущность–связь» [13], являющейся основой для унификации различных представлений данных (при этом следует отметить, что в статье [13] для обозначения связи между сущностями не используется термин «факт», а в ее библиографическом списке отсутствует ссылка на «Логико-философский трактат»).

Для единообразия определения понятия «факт» удобно использовать модификацию модели данных «сущность-связь» из той же статьи, называемую моделью множества сущностей. Ее отличительные особенности заключаются в том, что, во-первых, в ней всё трактуется как объекты (в том числе, например, цвет, в то время как в модели «сущность-связь» цвет обычно трактуется как «значение», а согласно «Логико-философскому трактату» «2.0251. Пространство, время и цвет (цветность) есть формы объектов») а, во-вторых, все связи в этой модели – бинарные. Связи между объектами в модели множества сущностей также рассматриваются как объекты, связанные, в свою очередь, с объектами – атрибутами связей.

Важно подчеркнуть, что создание фактографических систем подразумевает извлечение фактов не только непосредственно из текста документа, но и из его метаданных. Это следует, например, из традиционного понимания научно-информационного процесса [14], 2-й этап которого (аналитико-синтетическая переработка документальной информации) предусматривает как извлечение сведений о содержании документа (индексирование, аннотирование и т.п.), так и обработку его библиографических данных.

Заметим, что указание источника, из которого извлечен данный факт, в качестве одного из атрибутов факта, позволяет с той или иной степенью достоверности отделять «существующие» (в терминологии Витгенштейна) факты от «несуществующих». С этой целью на множестве источников может быть введена шкала их достоверности.

Таким образом, отношения и факты объявляются первичными, а вещи представляют собой пересечение, совокупность возможных отношений, то есть с вещью можно соотнести общую область «пересечения» множества фактов. Атомарный факт есть соединение (двух) объектов. Анализ фактов дает объекты или предметы. При этом по мере накопления фактов представление о вещи может меняться. Благодаря такой трактовке мира вещь выступает не как нечто данное, застывшее, вполне определенное, а как некоторая сущность с размытыми границами, и эти границы уточняются по мере выявления класса возможных для данной сущности отношений (фактов). Чтобы определить вещь, надо зафиксировать все факты (положительные – где может встречаться эта вещь и отрицательные, где не может) [11].

Отсюда вытекает, что функционирование интеллектуальной информационной системы основано на двух противоположных процессах: при пополнении ИС новыми сведениями происходит преобразование семантической информации в данные, однако непосредственно потребности

пользователя удовлетворяет обратный процесс – извлечение из данных нужной пользователю информации и знаний.

Однако всякий ли факт, содержащийся в тексте или метаданных документа, обрабатываемого ИнтС с целью извлечения из него фактов, представляет интерес с точки зрения создателей и пользователей данной ИнтС? Чтобы ответить на этот вопрос, формализуем введенное понятие факта подобно тому, как это было сделано в нашей работе [15] для терминов «информация», «знание», «тезаурус», «онтология». В этой работе, в частности, показано, что данные соответствуют синтаксическому уровню сообщения (в том числе документа), информация (в узком смысле!) – семантическому, а знания – прагматическому. Отсюда вытекает, что функционирование интеллектуальной информационной системы основано на двух противоположных процессах: при пополнении ИнтС новыми сведениями происходит преобразование семантической информации в данные, однако непосредственно потребности пользователя удовлетворяет обратный процесс – извлечение из данных нужной пользователю информации и знаний.

Следовательно, в качестве «первичного» факта рассматривается некоторая информация (как правило, семантическая; примеры возможных исключений приведены выше), но в справочно-информационный фонд ИнтС факт заносится в качестве совокупности элементов данных, описывающих сущности и связи между ними, что соответствует уже упоминавшемуся соотнесению данных и фактов из монографии [2].

Но какого рода информация может быть занесена в справочно-информационный фонд системы в виде данных? Ведь сами по себе данные не несут никакой информационной ценности без соответствующих моделей: например, А.Н.Колмогоров неоднократно отмечал, что данные представляют информационную ценность лишь тогда, когда они являются составной часть некоторой модели реального мира и связаны с другими данными [16, 17]. Таким образом, применение информационных технологий должно основываться на использовании различных моделей (феноменологических, информационных, математических и др.). Как подчеркивал А.А.Ляпунов (см., например, [18]): «нет модели – нет информации».

В качестве модели предметной области обычно выступает ее *онтология* (какая именно смысл мы вкладываем в это весьма широко трактуемое понятие – будет уточнено в следующем разделе).

Таким образом, при создании фактографических информационных систем разумно следующее понимание факта*:* **содержащаяся в тексте и метаданных документа совокупность связей между сущностями, описываемыми в онтологии информационной системы**.

Отсюда, в частности, вытекает следующее важное замечание: именно онтология фактографической системы определяет, что будет считаться фактом в рамках этой системы.

4 Модель онтологии фактографической системы

Поскольку «объектом сбора, хранения, поиска и выдачи в так называемых фактографических информационно-поисковых системах… могут быть лишь соответствующие тексты или документы, описывающие некоторые данные или факты, если под документом понимать… любой фрагмент такого текста» [10], постольку в роли онтологии – модели предметной области – может выступать та или иная модель интеллектуальной информационной системы, например предложенная нами в работе [19]. Эта модель, записанная в качестве модели предметной области, имеет вид

$$S = \langle K, M, M^j \langle K_i, K_{i'} \rangle \rangle,$$

где K – классы сущностей, M – множество используемых атрибутов сущностей, $M^j \langle K_i, K_{i'} \rangle$ – типы возможных связей между классами сущностей, когда сущность из класса $K_{i'}$ может входить в качестве значения атрибута M^j сущности из класса K_i. Тем самым, как отмечено выше, любая сущность s_i может быть представлена в виде

$$d_i = \langle m_i^{j,k} \rangle,$$

где $m_i^{j,k}$ – значения атрибутов сущности, k – количество значений (с учетом повторений) j-го атрибута в описании сущности.

При создании информационной системы сущности будут представлены в виде описывающих их документов, а атрибуты сущностей будут представлять собой элементы метаданных.

Предложенная модель онтологии полностью соответствует введенному нами пониманию факта, что делает ее наиболее пригодной для создания фактографической системы. Разумеется, пользуясь знаниями о предметной области, возможно и целесообразно накладывать различные ограничения (морфологические, синтаксические, семантические, структурно-текстовые) на характеристики сущностей, входящих в те или иные классы (подробно принципы установления ограничений описаны в [20]).

Отметим, что применительно к фактографическим информационным системам, создаваемым в рамках концепции Semantic Web, довольно близкий подход был предложен в работе [21]. Речь идет об использовании модели, в которой сущности внешнего мира представляются атрибутированными информационными единицами, а отношения между сущностями реализуются либо в виде прямых ссылок, либо в виде составных конструкций определенного вида, при этом спецификация такой модели воплощается в виде онтологии.

5 Заключение

В статье изложены модели документального и фактографического поиска в электронных библиотеках, работающих с документами достаточно произвольной структуры. Предложена модель классификации документов электронной библиотеки, основанная на использовании отношения толерантности, учитывающая возможное отсутствие априорно заданных классификаторов. Показано, что при создании фактографических систем целесообразно следующее понимание факта: содержащаяся в тексте и метаданных документа совокупность связей между сущностями, описываемыми в онтологии информационной системы. Предложена простейшая модель онтологии фактографической системы.

Литература

[1] В.Б. Барахнин, А.М. Федотов, О.А Федотова. Электронная библиотека по научному наследию как фактографическая система // Труды XV Всероссийской конф. «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» – RCDL'2013. – Ярославль, 2013. – С. 91–97.

[2] Ю.М. Арский, Р.С. Гиляревский, И.С. Туров, А.И. Черный. Инфосфера: Информационные структуры, системы и процессы в науке и обществе. – М.: ВИНИТИ, 1996.

[3] Ю.И. Шокин, А.М. Федотов, В.Б. Барахнин. Проблемы поиска информации. – Новосибирск: Наука, 2010.

[4] T. Berners-Lee, J. Hendler, O. Lassila. The Semantic Web // Scientific American. – 2001. Vol. 284(5). – P. 34–43.

[5] Ш.Р. Ранганатан. Классификация двоеточием. Основная классификация: пер. с англ. – М.: ГПТНБ СССР, 1970.

[6] А.М. Федотов, В.Б. Барахнин. Проблемы поиска информации: история и технологии // Вестник НГУ. Серия: Информационные технологии. – 2009. – Т. 7, вып. 2. – С. 3–17.

[7] Ю.А. Шрейдер. Равенство, сходство, порядок. – М.: Наука, 1971.

[8] T.A. Halpin. What is an elementary fact? // Proc. First NIAM-ISDM Conferenc. – Utrecht, 1993. – http://www.orm.net/pdf/elemfact.pdf

[9] M. Lezoche, A. Aubry, H. Panetto. Formal Fact-Oriented Model Transformations for Cooperative Information Systems Semantic Conceptualisation // Enterprise Information Systems. – Springer, 2012. – P. 117–131.

[10] А.И. Михайлов, А.И. Черный, Р.С. Гиляревский. Научные коммуникации и информатика. – М.: Наука, 1976.

[11] L. Wittgenstein. Logisch-Philosophische Abhandlung // Annalen der Naturphilosophie. Vol. XIV. Parts 3/4. Leipzig: Verlag Unesma, 1921. P. 185–262. – Рус. пер. Л. Витгенштейн. Логико-философский трактат. – М.: Изд-во иностранной лит., 1958.

[12] А.Ф. Грязнов. Витгенштейн // Новая философская энциклопедия. – М.: Мысль, 2000. Т.1. С. 406–408.

[13] P.P. Chen. The entity-relational model. Toward a unified view of data // ACM TODS. –1976. –№ 1. – P. 9–36.

[14] А.И. Михайлов, А.И. Черный, Р.С. Гиляревский. Основы информатики. – М.: Наука, 1968.

[15] В.Б. Барахнин, А.М. Федотов. Уточнение терминологии, используемой при описании интеллектуальных информационных систем, на основе семиотического подхода // Известия вузов. Проблемы полиграфии и издательского дела. – 2008. – № 6. – С. 73–81.

[16] А.Н. Колмогоров. Три подхода к определению понятия «количество информации» // Проблемы передачи информации. –1965. –Т. 1, вып. 1. – С. 3–11.

[17] А.Н. Колмогоров. Теория информации и теория алгоритмов. – М.: Наука, 1987.

[18] А.А. Ляпунов. О соотношении понятий материя, энергия и информация // Ляпунов А.А. Проблемы теоретической и прикладной кибернетики. – Новосибирск: Наука, 1980. – С. 320–323.

[19] В.Б. Барахнин, Ю.В. Леонова, А.М. Федотов. К вопросу о формулировке требований для построения информационных систем научно-организационной направленности // Вычислит. технологии. – 2006. – Т. 11. Спец.выпуск. – С. 52–58.

[20] Е.А. Сидорова. Онтологический подход к представлению знаний для задачи анализа текстовых ресурсов // Материалы Всероссийской конф. «Знания – Онтологии – Теории». – Новосибирск, 2007. – Т. 1. – С. 221–228.

[21] А.Г. Марчук. О распределенных фактографических системах // Труды X Всероссийской конф. «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» – RCDL'2008. – Дубна, 2008. – С. 93–102.

Models of Document and Factual Retrieval for Digital Libraries

V.B. Barakhnin, A.M. Fedotov

This paper considers the constructing of models of document and factual retrieval for digital libraries containing documents with arbitrary structure. A model of document classification based on the tolerance relation is proposed. This model takes into account the lack of predefined classifiers. It is shown that when creating factual information systems, the following definition of fact is advisable: this is a set of relationships between the entities contained in the document text and metadata and described in the ontology of the information system. A simplest model of ontology of factual system is proposed.