Построение моделей документального и фактографического поиска в электронных библиотеках

В.Б.Барахнин, А.М.Федотов

Институт вычислительных технологий СО РАН,
Новосибирский государственный университет

Постановка задачи

Электронные библиотеки (ЭБ), в т.ч. по научному наследию, выступают в качестве основы интеллектуальных систем (ИнтС), предназначенных не только для документального, но и для фактографического поиска, то есть позволяющих удовлетворять информационные потребности квалифицированного пользователя в соответствии со схемой “документ — факт — рассуждение”.

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. Эта задача особенно актуальна в электронных библиотеках, работающих с документами достаточно произвольной структуры, то есть без более или менее подробной стандартизации представления информации, например на основе словарей.

Модель классификации документов — понятие толерантности

Любой документ d_i каталога системы представляется как $d_i =< m_i^{j,k} >$. Рассмотрим подмножество метаданных M_C, определяющее набор классификационных признаков документов. Для фиксированного элемента метаданных M^j, где $M^j \in M_C$, множество документов разбивается на классы эквивалентности, соответствующие различным значениям этого элемента метаданных.

Будем считать два экземпляра сущностей **толерантными**, 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 $K_k^i \subset K_l^i$ 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.

Модель классификации документов — базис пространства толерантности

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

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.

Задание базовой структуры построения онтологии на основании многомерной классификации посредством ядер толерантности

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

Выделим подмножество элементов метаданных $M^* = \{M^{j_k}\}_{k=1}^{l}$, $M^{j_k} \subset M$, определяющее для данной предметной области важнейшие характеристики документов, при этом $M^{j_k} = \{m_i^{j_k}\}_{i=1}^{l_k}$. Тогда ядра толерантности, задающие базовую структуру, суть элементы декартова произведения $\mathbf{\Pi}M^* = M^{j_1} \times M^{j_2} \times \cdots \times M^{j_l}$. Поиск класса документов сводится к выбору соответствующего элемента $(m_{i_1}^{j_1}, m_{i_2}^{j_2}, \ldots, m_{i_l}^{j_l}) \in \mathbf{\Pi}M^*$, т. е. к отображению $S^* : \mathbf{\Pi}M^* \to D$, а предварительная классификация документов — к обратному отображению $C^* : D \to \mathbf{\Pi}M^*$.

1. 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”.

2. 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, а уже потом, исходя из результатов кластеризации, устанавливать подмножества множеств значений элементов метаданных, выступающих в качестве значений фасетов.

Уточнение понятия “факт”

При создании фактографических информационных систем разумно следующее понимание факта: **содержащаяся в тексте и метаданных документа совокупность связей между сущностями, описываемыми в онтологии информационной системы**. Это определение опирается на концепцию “Логико-философский трактат” Л.Витгенштейна, которая практически полностью воспроизводятся в модели данных “сущность-связь”

Удобно использовать модификацию модели “сущность-связь” , называемую моделью множества сущностей. Ее отличительные особенности заключаются в том, что, во-первых, в ней всё трактуется как объекты (в том числе, например, цвет, что соответствует “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, 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.

Из приведенного определения факта вытекает следующее важное замечание: именно онтология фактографической системы определяет, что будет считаться фактом в рамках этой системы

В роли онтологии — модели предметной области —- может выступать та или иная модель интеллектуальной информационной системы, например

$$S = <K, M, M^j<K_i, K_{i'}>>,$$

где K — классы сущностей, M — множество используемых атрибутов сущностей, $M^j<K_i, K_{i'}>$ — 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 $d_i = <m_i^{j,k}>$, 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.

Заключение

В докладе 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

Спасибо за внимание!