Некоторые аспекты индексации и поиска документов на основе вложенных многоуровневых структур

В.М.Куглер
Свердловская областная универсальная научная библиотека
им. В.Г. Белинского
valery@library.uraic.ru

Ключевые слова: коллекция документов, структура, запись, поле, информационный образ документа, поисковый образ документа.

Большие коллекции документов, будь то фонды библиотек, HTML страницы в Интернете, набор инструкций сложной инженерной системы и т.д., нуждаются в организации поиска. Для этой цели на практике создают формальные модели этих коллекций. Элементы модели являются образами элементов соответствующей коллекции, находятся с ними в одно – однозначном соответствии и представляют собой их формальное описание, включающее идентификацию, в частности адресную. Для книг создаются библиографические описания, к многотомным инженерным инструкциям – указатели и путеводители. Процесс создания формальных информационных образов реально существующих документов будем называть их индексацией.

Мы рассматриваем ситуацию, когда есть большая совокупность записей, имеющих некоторую структуру, возможно, многократной вложенности. Таким образом, запись состоит из полей, имеющих каждое имя и значение. Значение каждого поля может также состоять из подполей, имеющих имя и значение. Такой процесс вложенности может продолжаться неограниченно.

Будем придерживаться следующей нотации: поле записываеся в виде имени поля, после которого идет разделитель – двоеточие, за ним – значение поля в косых скобках. Поля отделяются точкой с запятой. Например:
< Автор:<Евтушенко>; Заглавие: <Нежность> >

С помощью структур можно описать широкий спектр объектов, дело в выработке договоренности между людьми, пользующимися совместно данным коллективом записей, о семантической значимости наименований полей.

Замечательным примером такой договоренности является Российский коммуникативный формат представления библиографических записей Rusmarc.

Мы можем, например, договориться о специальном значении наименований полей, когда создаем каталог ресурсов Интернет. Это «наименование», «атрибуты», «состав», «доступ», «коллекция». Сводный электронный каталог библиотек Свердловской области можно описать так:

< Наименование:< Сводный электронный каталог библиотек Свердловской области>;
Атрибуты:<Владелец:<УРГУ>; Владелец:<СОУНБ>; Владелец:<СГТУ> >;
Состав:<Коллекция:<Библиографическое описание:<Состав:<<автор>;<название>>>>;
Доступ: <поиск:<поисковое поле:<<все>;<автор>;<название>>;
результат:<библиографическое описание>>>
>

Альтернативой стандартам, где имена полей и структура определены на 100% являются системы со свободным выбором имен полей, значений и иерархии вложенности. Практическим приложением является, например, каталог ресурсов Интернет, где URL и структурированные описания может добавлять любой желающий. В такой совокупности структур для формулировки поискового образа должны быть средства фиксации неопределенности уровней вложенности.

Когда проблема оформилась в нашей голове и мы полагаем, что ее решение состоится благодаря документу, принадлежащему определенной коллекции, мы формулируем поисковый образ. Он отражает ту проблему, которая решается в нашей голове.

Кроме того он необходим, чтобы осуществить поиск в формальной модели коллекции, поэтому правила составления поискового образа основываются на языке формальной модели и также должны помочь пользователю достичь успеха в его поиске. Поскольку образы документов мы строим как иерархические структуры неограниченной сложности, следует подумать о том, что поисковый образ не должен быть сложным, а в случае неудачи, когда найдено ноль документов, должно быть понятным, как ослаблять выставленные требования.

Наиболее «мягким» вариантом будет поисковый запрос в виде одного и более слов, а правило соответствия структуры поисковому запросу: все его слова входят в структуру как имена или значения полей.

Далее поисковый запрос сам может быть структурой, и тогда «мягким» вариантом будет правило соответствия структуры запросу, когда его можно вложить в нее. Это означает: 1. Для каждого узла запроса существует узел структуры с таким же именем и это определяет одно – однозначное отображение запроса внутрь сравниваемой с ним структуры; 2. Если два узла запроса являются «соседями», то между соответствующими узлами в структуре имеется «путь». Трансформация этих двух пунктов в сторону полного совпадения структуры запроса и сравниваемой структуры дает варианты более жестких правил.

Например, нотация

<*>СОУНБ <*>каталог

может запрашивать структуры, где слова «СОУНБ» и «каталог» встречаются на разных произвольных уровнях вложенности.

Запрос:

<*><владелец:<СОУНБ>>

Выделяет все структуры, где фраза «владелец:<СОУНБ>» есть целиком, уровень глубины вхождения любой.

Выводы:

- Структуры позволяют совершенно точно описать информационный образ документа.
- В системах коллективного пользования документы могут индексироваться свободными структурами – ограничений на имена полей, значения, структурную вложенность нет.
- Поисковые запросы могут задавать структуру точно или с любой степенью варьирования положения слов относительно уровней вложенности.