Классификация документов в электронных библиотеках
Document Classification in Digital Libraries
Класифікація документів в електронних бібліотеках

О. В. Пескова
Московский государственный технический университет им. Н. Э. Баумана, Москва, Россия

Olga Peskova
Bauman Moscow State Technical University, Moscow, Russia

О. В. Пєскова
Московський державний технічний університет ім. М. Е. Баумана, Москва, Росія

Обсуждается классификация полнотекстовых документов в электронных библиотеках с точки зрения механизма реализации поисковых возможностей для читателей. Рассматриваются основные подходы к построению систем автоматической классификации полнотекстовых документов: категоризация текстов и кластеризация текстов.

Fulltext Document Classification in Digital Libraries is discussed regarding information retrieval purposes. Fundamental approaches to developing automatic text classification, such as text categorization and text clusterization, are considered.

Обговорюється класифікація повнотекстових документів в електронних бібліотеках з точки зору механізму реалізації пошукових можливостей для читачів. Розглядаються основні підходи до побудови систем автоматичної класифікації повнотекстових документів: категоризація текстів і кластеризація текстів.

В настоящее время во всём мире уже осознана необходимость комплектования документных фондов электронными документами, осознана также необходимость создания и развития электронных полнотекстовых библиотек, которые дают много преимуществ в обслуживании читателей. Одними из главных таких преимуществ являются поисковые возможности. Основными механизмами реализации поисковых возможностей электронной библиотеки являются информационный поиск по запросу пользователя и классификация документов фонда. В данном докладе мы сосредоточимся на обсуждении второго поискового механизма – классификации документов.

Классификация полнотекстовых документов как механизм реализации поиска может применяться следующим образом:

- во-первых, когда неопытный в какой-то определённой предметной области читатель желает найти необходимую ему литературу, он испытывает трудности со стандартным поиском документов по ключевым словам, тогда ему на помощь приходит система навигации по классификационной схеме, которая позволяет углубляться в выбранную предметную область, и, таким образом, читатель легко находит те документы, которые соответствуют его требованиям;
- во-вторых, имеющиеся в электронной библиотеке поисковые системы могут использовать информацию о классификации документов с целью изменения ширины поисковой области, таким образом сокращая число нерелевантных документов в результатах поиска;
- и, в-третьих, механизм классификации документов может быть применён для группировки списков документов, полученных в результате работы поисковых систем. Вспомним, как часто многие из нас при работе с поисковыми системами сталкиваются с проблемой отбора нужных документов среди списка документов, который выдаётся в ответ на заданный запрос. Этот список может содержать тысячи релевантных документов, очевидно, что читатель не сможет просмотреть их все, и возможно, так и не найдёт требуемую ему информацию; облегчить ему эту задачу можно, если результаты работы поисковой машины представлять в виде некоторых тематических групп, т. е. разбить на классы релевантные запросу документы.

В первом варианте использования классификации можно говорить о классификации документов как о самостоятельном поисковом механизме, а во втором и третьем классификация выступает как средство повышения качества и удобства поисковых систем.

Очевидно, что классификация документов способствует повышению качества обслуживания читателей, однако, при современном быстром темпе развития информационных массивов, нетрудно представить, каким чрезмерно трудоёмким процессом будет классификация всего фонда электронных документов вручную. Облегчить эту задачу сотрудникам документного фонда могут программные средства, способные автоматически выполнять классификационные действия над документами. В действительности, в настоящее время уже стало возможным воплощение идеи автоматической классификации документов, поскольку, во-первых, речь идёт о полнотекстовых документах, которые могут быть представлены в виде пригодном для автоматического анализа с помощью программных средств, во-вторых, к настоящему моменту в научном сообществе накопился достаточно большой опыт исследования и разработки таких систем, причём интерес к данной проблеме среди исследователей не только не угасает, но в последние два десятилетия является повышенным, что в первую очередь вызвано скачком в развитии аппаратной базы, которая стала пригодной для тестирования разработанных ранее математических методов автоматической классификации.

Методы автоматической классификации полнотекстовых документов можно разделить на следующие две группы:

1) Автоматическая классификация полнотекстовых документов с обучением, или категоризация документов: документы классифицируются по предопределенной классификационной схеме (рубрикатору) на основании знаний о том, какими признаками должны обладать документы, чтобы относиться к той или иной рубрике, т. е. для каждой рубрики имеется некоторое множество документов-образцов. Иначе говоря, методы категоризации полнотекстовых документов возможны для использования с целью классификации фонда электронных документов, если:

– во-первых, уже сделан выбор классификационной схемы, например, принято решение в фонде электронных документов применять традиционные библиотечные классификационные схемы (УДК, ГРНТИ, ББК);

– во-вторых, есть возможность сформировать обучающее множество документов, т. е. для каждой рубрики (индекса) классификационной схемы сформировать, например, с помощью экспертов, небольшое множество документов, соответствующих данной рубрике (индексу).

Для реализации программной системы категоризации документов разработчикам необходимо на основе анализа существующих математических методов категоризации [5] совершить выбор наиболее подходящего алгоритма и соответствующего ему способа формирования представления полнотекстового документа. В результате разработанная программная система будет автоматически классифицировать по заданной классификационной схеме новые поступления в электронную библиотеку.

2) Автоматическая классификация полнотекстовых документов без обучения, или кластеризация документов: документы классифицируются в условиях отсутствия предопределенной классификационной схемы и множества документов-образцов, т. е. группируются некоторым образом исходя из анализа тематического сходства между документами всей коллекции. Иначе говоря, методы кластеризации полнотекстовых документов возможны для использования с целью классификации фонда электронных документов, если:

– во-первых, отсутствуют строгие ограничения на выбор классификационной схемы, например, принято решение использовать собственный, сформированный для данной коллекции предметный рубрикатор;

– во-вторых, в качестве итоговой классификации коллекции документов допустима просто группировка документов в соответствии с их тематическим сходством. Такая группировка документов может быть как иерархической, так и простым набором групп. Классификационной схемой в таком случае будут выступать автоматически сформированные названия групп и отношения между ними, названия в общем случае представляют собой списки ключевых слов для каждой группы (кластера) документов.

Для реализации программной системы кластеризации документов разработчикам необходимо на основе анализа существующих математических методов кластеризации [6] совершить выбор

наиболее подходящего алгоритма и соответствующего ему способа формирования представления полнотекстового документа. В результате разработанная программная система, во-первых, позволит существенно сократить трудоёмкость процессов как формирования классификационной схемы, так и классификации по ней документов, во-вторых, раскроет внутреннюю скрытую тематическую структуру электронного документного фонда.

Благодаря разнообразию математических методов, положенных в основу известных алгоритмов автоматической классификации полнотекстовых документов, есть возможность сделать успешный выбор оптимального решения задачи автоматической классификации в каждой конкретной электронной библиотеке, что позволит предоставить читателям удобный инструмент поиска и навигации по коллекции документов, при этом сократив трудоёмкость реализации данного инструмента за счёт внедрения нового технологичного решения.

Литература

1. Sebastiani F. Machine Learning in Automated Text Categorization // ACM Computing Surveys. – 2002. – Vol. 34, No. 1.
2. Manning D., Schutze H. Foundations of statistical natural language processing // The MIT Press. – 2003.
3. Прикладная статистика: Классификация и снижение размерности: Справ. изд. / С. А. Айвазян, В. М. Бухштабер, И. С. Енюков, Л. Д. Мешалкин; Под. ред. С. А. Айвазяна. – М.: Финансы и статистика, 1989. – 607с.: ил.
4. A. K. Jain, M. N. Murty, P. J. Flynn Data Clustering: A Review \\ ACM Computing Surveys. – Vol. 31, No. 3. – 1999.
5. Пескова О. В. Методы автоматической классификации текстовых электронных документов // Научно-техническая информация. Сер. 2. – 2006. – № 3. – С. 13–20.
6. Пескова О. В. Методы автоматической классификации электронных текстовых документов без обучения // Научно-техническая информация. Сер. 2. – 2006. – № 12. – С. 21–32.