К ПРОБЛЕМЕ ПОСТРОЕНИЯ МОДЕЛИ ПОИСКОВОГО ОБРАЗА ДОКУМЕНТА

Леонова Ю.В., Федотов А.М.
Институт вычислительных технологий СО РАН, Новосибирск
e-mail: juli@ict.nsc.ru, fedotov@ict.nsc.ru

Аннотация
В докладе рассматривается подход к решению задачи автоматизации классификации документов. Качество классификации напрямую зависит от содержания моделей, по которым построена поисковая система. Авторами рассматривается модель представления поискового образа документа.

Важной задачей информационного поиска является отнесение документа к одной из нескольких категорий на основании семантического содержания документа, называемой задачей классификации. Классификация может осуществляться полностью вручную, либо автоматически с помощью заданного набора решающих правил.

Постановка задачи

Имеется множество категорий (классов) $\mathfrak{C} = \{c_1, \dots, c_{|\mathfrak{C}|}\}$.

Имеется множество документов $\mathfrak{D} = \{d_1, \dots, d_{|\mathfrak{D}|}\}$.

Имеется неизвестная целевая функция – отображение $\Phi: \mathfrak{C} \times \mathfrak{D} \to [0,1]$, которая классифицирует документ и выдает степень принадлежности [1,2]. Требуется определить данную целевую функцию.

Необходимо реализовать процедуру автоматического отнесения произвольного документа к одному или нескольким классам – каждый поступающий в систему документ должен автоматически соотноситься с одним или несколькими классами, либо ни с каким классом. Процесс классификации должен быть именно автоматическим: классификация происходит без помощи специально обученных экспертов, которые могут определить, попадает тот или иной документ в определенный раздел, никакого подбора правил вручную.

Возможны два способа задания классов:

- эмпирически;
- аксиоматически.

При эмпирическом способе задание множества классов осуществляется экспертом на основе непосредственного сравнения общих свойств некоторого класса документов.

Аксиоматическое задание множества классов осуществляется на основе опосредованного анализа некоторого класса документов при помощи ранее выработанных условий или правил, которым должен удовлетворять документ.

На практике используется комбинация этих способов – над частью документов задается аксиоматическое отнесение к классам, а для документов, загружаемых пользователями, – эмпирическое.

Имеется множество способов определения целевой функции. Наиболее популярный способ – задание меры сходства на множестве документов. Для каждого класса задается документ-центроид, значение функции – мера сходства между поисковым образом документа и образом центроида класса [9].

Каждый документ описывается набором своих характеристик, называемых *признаками*. Признаки могут быть числовыми или нечисловыми, простыми и сложными.

Определение признаков – формирование поискового образа документа (ПОД) – вектор характерных признаков документа, используемый в дальнейшем для принятия решений по работе с документом. ПОД представляет собой многомерный вектор в пространстве признаков документа, характеризующих смысловое содержание исходного документа.

Признаковое пространство

Признаком называется отображение $f: \mathfrak{D} \to \mathbb{D}_f$, где $\mathbb{D}_f$ – множество допустимых значений признака. Если заданы признаки $f_1, \dots f_n$, то вектор $\mathbf{d} = (f_1(d), \dots, f_n(d))$ называется *признаковым описанием* документа $d \in \mathfrak{D}$. Признаковые описания допустимо отождествлять с самими документами. При этом множество $\mathfrak{D} = \mathbb{D}_{f_1} \times \dots \times \mathbb{D}_{f_n}$ называют *признаковым пространством*.

В зависимости от множества $\mathbb{D}_f$ признаки делятся на следующие типы:

- *бинарный* признак: $\mathbb{D}_f = \{0,1\}$ (пол человека);
- *ранжированный* признак: $\mathbb{D}_f \in [0,1]$ (мера близости);
- *номинальный* признак: $\mathbb{D}_f$ – конечное множество (место проживания, профессия, работодатель);
- *порядковый* признак: $\mathbb{D}_f$ – конечное упорядоченное множество (образование, занимаемая должность);
- *количественный* признак: $\mathbb{D}_f$ – множество действительных чисел (год издания, возраст, стаж работы);
- *иерархический* признак: $\mathbb{D}_f$ – конечное частично-упорядоченное множество (термины в тезаурусе).

Для некоторых типов признаков (например, номинальных) при наличии словаря значения для соответствующего признака документа выбираются из строго фиксированного множества слов, ограниченного набором тщательно подобранных терминов. Это может очень серьезно улучшить возможности автоматической классификации, а также повысить качество результатов поиска, поскольку поисковые системы хороши в побуквенном сравнении слов, но чувствуют себя намного хуже, когда описание термина сделано в «человеческом» стиле, с синонимами, контекстом и т.д. Однако, как правило, поисковые признаки документов не всегда точны, поскольку часто задаются авторами без терминологического контроля, например, УДК может быть неправильно задан, ключевые слова могут не в полной мере отражать содержание документа и т.п.

Для каждой пары <документ, категория> определяется степень принадлежности документа категории, причем эта степень является числом, лежащим в диапазоне от 0 до 1 (чем больше степень принадлежности, тем больше документ подходит этой категории).

В нашем случае используется следующий порядок обработки документов. Имеется набор категорий, и поступает новый документ, для которого нужно определить список подходящих ему категорий. Если документ не попадает ни в одну категорию, то он отбрасывается.

На множестве категорий могут быть заданы теоретико-множественные отношения. Например, множества документов, составляющих категории, могут пересекаться или не пересекаться, т.е. один и тот же документ может принадлежать нескольким категориям. Могут быть заданы составные категории – категория может быть подмножеством другой категории и т.п.

Схемы представления поисковых признаков документа

Можно выделить 3 типа признаков:

- Базовые (Dublin Core)
- Основные (ГОСТ)
- Дополнительные
- Сложные (вычислимые)

Рассмотрим существующие схемы представления поисковых признаков документа.

Базовые признаки присутствуют в описании каждого документа, задаются стандартом Dublin Core. Набор элементов метаданных Дублинского ядра (Dublin Core Metadata Element Set; DCMES) состоит из 15 элементов метаданных [3,4]:

1. Title – название;
2. Creator – создатель;
3. Subject – тема;
4. Description – описание;
5. Publisher – издатель;
6. Contributor – внёсший вклад;
7. Date – дата;
8. Type – тип;
9. Format – формат документа;
10. Identifier – идентификатор;
11. Source – источник;
12. Language – язык;
13. Relation – отношения;
14. Coverage – покрытие;
15. Rights – авторские права.

Этот набор включает частично элементы, которые имеют одинаковые значения в совершенно разных предметных областях, например, относящиеся к созданию, именованию и определению предметной области документов. В то же время другие, вероятно, выходят за рамки общепринятых, в частности, темпоральные (временные) и пространственные характеристики (элемент Coverage), а также элементы, ответственные за описание прав интеллектуальной собственности.

Каждый базовый элемент Дублинского ядра (Creator – «создатель», Title – «название» и т.д.) обладает дополнительными атрибутами (квалификаторами), при желании позволяющими более точно описать соответствующий блок информации. Таким образом, информация в DC в общем случае получается «двуслойной» и состоит из двух уровней – базового уровня (информация в самих элементах) и дополнительного уровня (информация в квалификаторах элементов).

В DC также обеспечена возможность создания продвинутых («профессиональных») метаописаний документов, включающих разнообразную, подчас специализированную, информацию о предмете описания. Для этого в DC включен механизм расширения набора элементов. Предполагается, что пользователи могут создавать и развивать свои собственные наборы элементов, настроенные на нужды их сообществ и предметных областей. Однако, поскольку ядром этих описаний служит DC, то даже очень специализированные описания будут иметь общую часть, понимаемую всеми.

В зависимости от типа документа (элемент Type) могут использоваться различные схемы расширения признаков. Например, для документа типа «публикация» для расширения набора поисковых признаков могут использоваться библиографические стандарты и ГОСТы. В библиотечном деле составление библиографического описания (совокупности библиографических сведений о документе) выполняется по правилам, установленным ГОСТ 7.1-841 «Библиографическое описание документа» [5]. В настоящее время насчитывается около 300 поисковых признаков. Поисковые признаки разделяются на два типа: основные и дополнительные.

Основные поисковые признаки описывают наиболее полные сведения, необходимые для идентификации и поиска издания – название издания, автор. Некоторая часть этих признаков дублирует схему Dublin Core. Дополнительные признаки используются для расширения поиска документа по элементам. Дополнительные признаки содержат указание на классификатор, жанр, язык документа, формат и место хранения и т.п.
Заметим, что для поисковой системы большинство дополнительных характеристик несущественны, в отличие от характеристик основного типа. Для реализации эффективного поиска основной интерес представляют классификационные дополнительные признаки, содержащие коды библиотечно-библиографических классификаторов.
Основные библиотечно-библиографические классификаторы:

- система библиотечно-библиографической классификации (ББК);
- универсальная десятичная классификация (УДК);
- десятичная классификация М. Дьюи (ДКД);
- единая классификация литературы для книгоиздания (ЕКЛ).

В основу библиотечно-библиографических классификаций положен тематический принцип.

Возможны также сложные признаки, основанные на использовании вычислимых характеристик текстов, например авторство текста документа. В общем случае текст отображается в вектор вычисленных для него параметров, каждый из которых объективно характеризует некоторый набор особенностей текста. Таким образом, текст графически отображается в некоторую точку n-мерного пространства. При такой формализации автор также может быть представлен в виде аналогичного вектора параметров – этим вектором будет вектор текстов, написанных данным автором.

Другими сложными признаками могут быть:

- пристатейные списки литературы;
- список персон, встречающихся в тексте документа;
- список географических объектов, встречающихся в тексте документа;
- список ключевых слов из тезаурусов и т.д.

Шкалирование и метризация пространства признаков

Задача отнесения неизвестного документа к одному из априорно заданных классов таких документов аналогична задаче распознавания образов [6], суть которой состоит в том, что при распознавании требуется, используя априорную информацию о принадлежности известных документов к соответствующим классам, указать принадлежность нового неизвестного документа, описание которого в общем случае не совпадает ни с описаниями классов, ни с описаниями ни одного из известных документов; если это невозможно, то следует отказ от распознавания.

Для применения алгоритмов распознавания и классификации необходимо метризовать пространство признаков. В целях обеспечения однородности и метризуемости пространства признаков необходимо шкалировать это пространство. Другими словами, для сравнения документов необходимо ввести формальную меру сходства (различия) объектов, в терминах которой и будут сравниваться документы между собой [7], а точнее – будут сравниваться поисковые образы документов.

Среди шкал обычно выделяют 3 основных типа:

- количественные, когда признаки измеряется в некоторой шкале, например, длины, веса, скорости и т.п. и числа, которыми выражаются значения количественных признаков, показывают на или во сколько раз различаются объекты по данному признаку и допускают любые арифметические преобразования с ними – количественная шкала;
- качественные, когда проявления признаков можно естественным образом упорядочить по их значениям (сила ветра в баллах, оценки на экзаменах или в спортивных соревнованиях и т.п.). При этом градации качественных признаков обычно выражаются упорядоченным рядом слов (плохо, удовлетворительно, хорошо, отлично) или целыми числами (шкала твердости минералов), обычно увеличивающимися с возрастанием степени проявления соответствующего признака. Однако такие числа (точнее цифры-символы) нельзя суммировать или умножать, поскольку эти цифры выражает только место (имя) градации в их последовательности – шкала порядков;
- классификационные, когда различные проявления признака можно упорядочить только на уровне наличия или отсутствия, т.е. на уровне классов объектов, например, пол, место жительства, профессия и т.п. – шкала наименований

Рассмотрим задачу метризации и шкалирования более конкретно на примере информационной системы “Модели изменения биосферы на основе баланса углерода (по натурным и спутниковым данным и с учетом вклада бореальных экосистем)” [8]. Для метризации пространства признаков необходимо определить наиболее значимые признаки, на основе которых строятся метрики. Качественные и классификационные шкалы строятся на основе классификационных признаков (атрибутивных) – УДК, ББК и т.п. Данные признаки, как правило, имеют большой вес. Применение количественных шкал для построения метрик возможно на основе эмпирических данных.

1) Распределение ключевых терминов в тексте документа.
2) Распределение фамилий персон, упоминаемых в документах. Часто коллектив авторов, работающий в определенной области, в текстах статьи ссылается на других персон, работающих в данной области. Поэтому данный признак может использоваться в качестве классификационного.
3) Распределение территорий, упоминаемых в документах. Этот признак, прежде всего, важен для задач, имеющих экологическую тематику. Упоминание территорий в документах привязывает работу к географическому местоположению и таким образом также может являться классификационным признаком.

На основании полученных распределений эмпирически строятся метрики, т.е. определение расстояния между документами для этих признаков при использовании алгоритмов распознавания возможно только экспериментально.

Заключение

Предложенная технология может быть применена для дальнейшего расширения электронной библиотеки.

ЛИТЕРАТУРА

[1]. Christopher D. Manning, Prabhakar Raghavan, Hinrich Schütze An Introduction to Information Retrieval Draft // Online edition. Cambridge University Press. - 2009. - 544 p.
[2]. Ю. Лифшиц Лекция № 6 по классификации текстов курса «Современные задачи теоретической информатики»//[http://yury.name/modern/]
[3]. А.В. Манцивода Система метаописаний Dublin Core// TeaCode [http://teacode.com/concept/eor/dc.html]
[4]. The Dublin Core® Metadata Initiative // [http://dublincore.org/]
[5]. ГОСТ 7.1-2003 Библиографическая запись. Библиографическое описание. Общие требования и правила составления
[6]. Н.Н. Малюков Лекции по прикладному анализу данных. Распознавание образов // [http://prand.ru/content/ raspoznavanie-obrazov]
[7]. Н.Н. Малюков Лекции по прикладному анализу данных. Стандартизация и шкалирование данных // [http://prand.ru/content/standartizatsiya-i-shkalirovanie-dannykh]
[8]. Информационная система “Модели изменения биосферы на основе баланса углерода (по натурным и спутниковым данным и с учетом вклада бореальных экосистем)” // [http://www.sbras.ru/win/elbib/data/show_page.dhtml?2+330]
[9]. Шокин Ю.И. Проблемы поиска информации / Ю.И. Шокин, А.М. Федотов, В.Б. Барахнин – Новосибирск: Наука, 2010 – 196 с.