

ФАСЕТНАЯ ОРГАНИЗАЦИЯ ИНТЕРНЕТ-КАТАЛОГА И АВТОМАТИЧЕСКАЯ ЖАНРОВАЯ КЛАССИФИКАЦИЯ ДОКУМЕНТОВ

П.И. Браславский

ИМаш УрО РАН

pb@dpt.ustu.ru

Е.А. Вовк

ООО «Яндекс»

lenka@yandex-team.ru

М.Ю. Маслов

ООО «Яндекс»

maslov@yandex-team.ru

Ключевые слова: интернет-каталог, фасетная классификация, автоматическая классификация документов, стилистическая классификация, жанровая классификация, дискриминантный анализ

Объем и разнообразие ресурсов интернета приводят к сложности иерархических схем классификации универсальных интернет-каталогов. В качестве возможного решения этой проблемы предлагается подход на основе фасетной классификации (ФК). Каталог «Яндекс», построенный с использованием ФК, имеет фасет «жанр» наряду с другими. В работе описывается разработка автоматической процедуры классификации документов по жанрам; приводятся и обсуждаются результаты тестирования процедуры. Делается вывод об эффективности полученной процедуры, формулируются предложения по дальнейшему развитию и применению метода.

1. Введение

Каталоги ресурсов – популярное средство поиска информации в интернете. Однако количество и разнообразие ресурсов Сети приводят к тому, что схема иерархической классификации универсальных интернет-каталогов становится громоздкой и труднообозримой. В качестве одного из возможных решений этой проблемы можно предложить *фасетную классификацию* (ФК).

ФК – это совокупность нескольких независимых классификаций, осуществляемых одновременно по различным основаниям [4].

Во-первых, ФК предоставляет пользователю многоаспектное описание ресурсов интернета, т.е. учитывается то, что поисковый запрос описывается не только тематическим пространством возможного ответа. Пользователь может также ориентироваться как на некоторый тип знания, интересующий его в связи с его информационной потребностью, так и на смысловой контекст ответа.

Во-вторых, на основании ФК возможно построение семантических типологий ресурсов – выделение внутри классов тех категорий, которым соответствует самостоятельный смысл материалов.

В-третьих, ФК разгружает тематическую структуру каталога, что позволяет перейти к более обозримой структуре без потери детальности классификации.

Однако при использовании ФК задача редакторов каталога усложняется: во-первых, им приходится «позиционировать» ресурс в пространстве большей размерности; во-вторых,

немалые трудности представляет сохранение общности трактовки признаков в разных контекстах. Поэтому в каталогах на основе ФК методы автоматической классификации ресурсов приобретают большое значение.

Каталог «Яндекс», построенный с использованием ФК, имеет фасет Жанр наряду с другими. В статье описывается разработка автоматической процедуры жанровой классификации документов, приводятся и обсуждаются результаты тестирования процедуры. Делается вывод об эффективности полученной процедуры, формулируются предложения по дальнейшему развитию и применению метода.

2. Фасетная организация каталога «Яндекс»

Каталог «Яндекс», построенный с использованием ФК, был запущен в октябре 2000 года по адресу <http://www.yandex.ru>. В марте 2002 года состоялся запуск его новой редакции по адресу <http://yaca.yandex.ru>. Тематическая структура каталога «Яндекс» более обзрима по сравнению с традиционными универсальными каталогами российской части интернета: каталог «Яндекс» содержит порядка 500 тематических рубрик вместо 5-10 тыс. для традиционных.

Значения фасетов для ресурсов интернета проставляются вручную редакторами при описании ресурсов в каталоге «Яндекс». На март 2002 г. в каталоге содержалось описание более 43 тыс. *ресурсов*, т.е. сайтов или их структурных разделов.

Информация о значениях фасетов также передается из каталога в поисковую систему Яндекс. Благодаря этому пользователи могут уточнить результат поиска по теме или региону. Ясно, что описать вручную каждый из 60 млн. документов российского интернета, а именно столько проиндексировал Яндекс к марту 2002 года, невозможно. Поэтому значения фасетов документам приписываются в соответствии с принципом *наследования* от вышележащего домена или директории. Благодаря этому около 70% документов, известных поисковой системе Яндекс (порядка 40 млн.), имеют хотя бы один унаследованный признак из каталога.

Основные фасеты, используемые в каталоге, это: Тема, Регион, Жанр, Источник информации, Адресат информации, Сектор экономики.

Тема имеет порядка 600 значений и описывает предметную область интернет-ресурса. Значение признака определяется содержанием текстов, сферой деятельности представляемой организации или областью применения предлагаемого продукта. Тематическая структура представляет собой иерархический классификатор с непересекающимися категориями.

Регион определяет принадлежность ресурса к одному из 230 географических областей. Принадлежность ресурса к региону может определяться несколькими показателями: географическим расположением представляемого объекта (например, местоположение торговой фирмы), сферой управления и влияния (регион действия политической партии), потенциальной аудиторией информации (целевая аудитория новостного издания) или информационным содержанием ресурса (справочник фирм города или история страны).

Источник информации имеет пять значений: Официальный, СМИ, Неформальный, Персональный Анонимный. С каждым из этих пяти типов источника информации связаны свои особенности подачи предоставляемых сведений – оперативность, достоверность, полнота, уникальность, объективность и т.п.

Адресат информации имеет четыре значения: Партнеры, Инвесторы, Потребители, Коллеги. Признак определяет аудиторию, для которой материалы могут представлять интерес – обычные покупатели, оптовики или фирмы-потребители, инвесторы или специалисты.

Сектор экономики имеет три значения: Государственный, Коммерческий, Некоммерческий. Признак присваивается сайтам организаций или частных предпринимателей и указывает на их экономико-правовой статус.

Значение фасета **Жанр** в каталоге «Яндекс» определяет принадлежность ресурса к одному из шести классов:

- художественная литература (*ХудЛит*);
- научно-техническая литература (*НаучТех*);
- научно-популярная литература (*НаучПоп*);
- нормативные документы (*НормДок*);
- советы;
- публицистика (*Публиц*).

Жанр описывает особенности содержания текстовых материалов. Значение фасета указывает на тип излагаемого знания (житейский, донаучный, научный, художественный [6]) и на тип запросов, которым ресурс может соответствовать.

Признак «Художественная литература» описывает ресурсы, содержащие литературные произведения. Разница между признаками «Научно-техническая» и «Научно-популярная» литература – в специализированности текста, в необходимой степени подготовленности читателя и в способе изложения материалов. К научно-технической литературе относятся публикации для специалистов, техническая специальная литература. К научно-популярной – упрощенное изложение для начинающих и интересующихся в форме «просто о сложном». В категорию «Нормативные документы» попадают любые нормирующие тексты (ГОСТы, законы, распоряжения, правила), а также формы документов. Категорией «Советы» выделяются житейские рекомендации и рецепты, жизненный опыт, советы специалистов и другие правила практического свойства. К «Публицистике» относятся материалы общеполитического содержания, характерные для новостных и общественно-аналитических изданий.

В отличие от других фасетов, Жанр (наряду с фасетом Тема) достаточно хорошо определяется содержанием документов ресурса, в частности, их лексико-грамматическими характеристиками. Это открывает возможность для определения его значений в автоматическом или полуавтоматическом режиме. И хотя жанровая (стилистическая) классификация не так распространена в информационном поиске, как тематическая, опыт разработки этих методов существует. Так, например, в [7] описаны эксперименты по автоматической классификации корпуса английских текстов на 15 жанровых категорий.

3. Методика построения жанровой классификации

На момент начала разработки описываемого алгоритма структура жанров была несколько иной, чем это описано выше. Так, *Публицистика* не была выделена в самостоятельный класс, а *Научно-Техническая* и *Научно-популярная* литература не были разделены между собой. Поэтому при реализации макетного варианта процедуры мы рассматривали четыре жанра из шести, исключив *Публицистику* и считая *Научно-популярную* литературу частью *Научно-технической*. Таким образом, наша задача сводилась к построению автоматической процедуры, которая относит текстовый документ к одному из четырех жанров или отклоняет его. Процедура в свою очередь состоит из следующих этапов: вычисление параметров документа, собственно классификация, оценка достоверности классификации.

В соответствии с подходом, изложенным в [1], первоначально нами был сформирован первичный набор параметров классификации (более 30 параметров). Определяющими критериями при составлении первичного набора были простота вычисления параметров и их потенциальная значимость для жанровой классификации. Все параметры можно разделить на четыре группы:

- 1) Морфологические: спектр документа по частям речи, спектр глагольных форм и т.п.
- 2) Лексические: доля слов из заданных списков в документе.
- 3) Синтаксические (уровень словосочетания): доля предложений с цепочками существительных в родительном падеже, доля предложений с конструкциями ‘{можно|нужно} + инфинитив’ и др.
- 4) Формальные: средняя длина слова в буквах, средняя длина предложения в словах, доля предложений с экспрессивной пунктуацией (хотя первый из этих параметров можно отнести к лексике, а два других – к синтаксису, мы предпочитаем выделять их в отдельную группу).

Для того, чтобы исключить элементы оформления документов, все параметры вычислялись на основе 100 предложений (или меньшего числа, если документ короткий) в средней части документа. Все морфологические и синтаксические параметры, использующие морфологическую информацию, вычислялись с помощью программы **mystem** (разработчики И. Сегалович и В. Титов, программу можно получить на сайте <http://corpora.narod.ru>). Грамматическая омонимия не учитывалась: использовался первый вариант нормальной (словарной) формы, которую возвращает программа. Слова из латинских букв, а также предложения из таких слов не учитывались.

Для построения классификации мы использовали методы линейного дискриминантного анализа [2, 3], а в качестве дополнительных оценок классификации – методы, основанные на обобщенных метриках расстояния [2,5].

На начальном этапе была сформирована обучающая выборка (ОВ), которая стала основой построения классификации. ОВ состояла из 285 документов, распределение документов по жанрам: *ХудЛит* – 77, *НаучТех* – 48, *НормДок* – 94, *Советы* – 66.

4. Разработка процедуры жанровой классификации

Хорошим исходным пунктом построения классификации может служить наглядное представление данных, полученное с помощью процедур канонического анализа [3]. Суть метода состоит в линейном преобразовании исходного пространства параметров таким образом, чтобы первые координаты нового пространства наилучшим образом отображали взаимное расположение объектов. Диаграммы рассеяния документов ОВ в координатах первых трех канонических направлений (рис. 1, 2) дают представление о геометрической структуре четырех классов. С помощью трех канонических направлений мы можем удовлетворительно различить все три класса, несмотря на отсутствие четких границ между ними и, например, сильный разброс документов класса *НаучТех*.

После построения серии пробных классификаций мы перешли к сокращению размерности пространства классификации. Целью было составление минимального набора параметров классификации, который бы классифицировал ОВ не хуже, чем первичный набор. Для оценки дискриминационной значимости отдельного параметра используется значение F-статистики [2, 3]. В результате мы получили линейную дискриминантную функцию, которая относит текст к одному из четырех жанров (*ХудЛит*, *НаучТех*, *НормДок*, *Советы*) на основе семи параметров:

- 1) доля глаголов в личной форме;
- 2) доля наречий;
- 3) доля слов из списка «НаучТех»;
- 4) доля слов из списка «НормДок»;
- 5) доля слов из списка «Советы»;
- 6) средняя длина русского слова;

7) доля предложений с конструкцией '{можно|нужно} + инфинитив'.

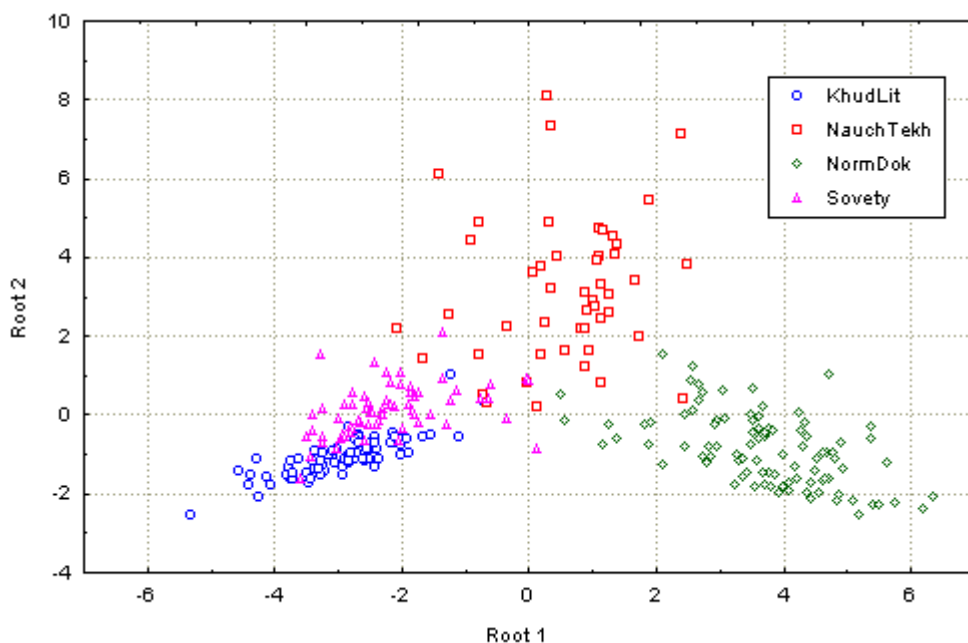


Рис. 1. Обучающая выборка в координатах первого (Root 1) и второго (Root 2) канонических направлений

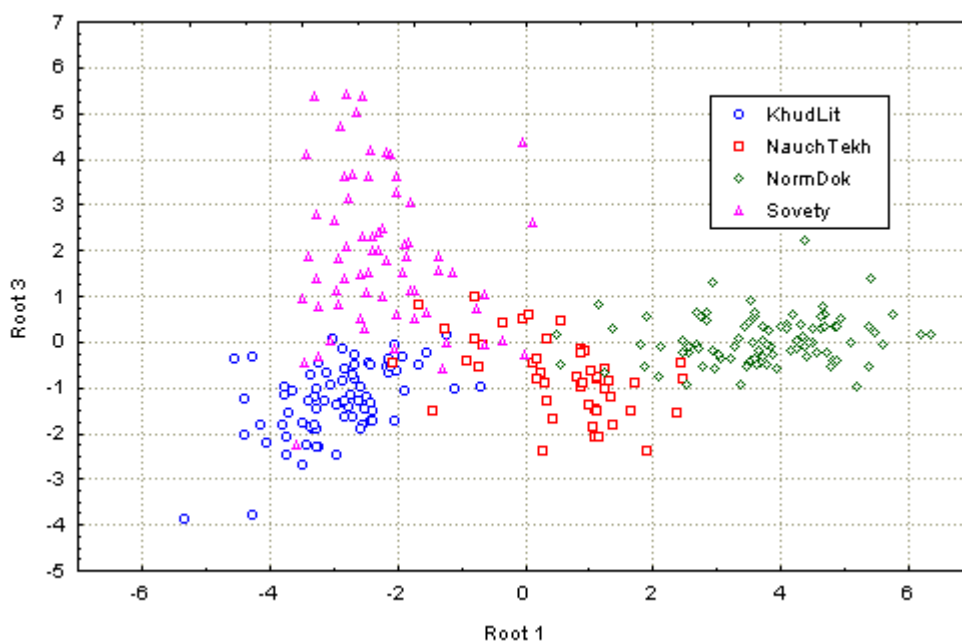


Рис. 2. Обучающая выборка в координатах первого (Root 1) и третьего (Root 3) канонических направлений

Для оценки того, насколько успешно полученная процедура выявляет структуру классов, была проведена классификация обучающей выборки (ОВ). Результаты этой пробной классификации сведены в таблице 1. Столбцы показывают, как были классифицированы жанры, представленные в ОВ. Например, из 77 документов *ХудЛит* 76 были классифицированы правильно, а один ошибочно отнесен к *Советам*. В последней строке – полнота классификации по жанрам. В строках – жанры, сформированные в результате классификации. Например, в *НаучТех* наряду с 42 «правильными» документами попали два из *НормДок* и один из *Советов*. В последнем столбце – точность классификации по жанрам. Точность и полнота совпадают (0,944), когда мы рассматриваем ОВ в целом.

Таблица 1. Матрица классификации обучающей выборки

	ХудЛит	НаучТех	НормДок	Советы	Все-авто	Точность
ХудЛит	76	1	.	6	83	0.916
НаучТех	.	42	2	1	45	0.933
НормДок	.	1	92	.	93	0.989
Советы	1	4	.	59	64	0.922
Все-ОВ	77	48	94	66	285	0.944
Полнота	0.987	0.875	0.979	0.894	0.944	

Ясно, что четыре выбранных жанра не покрывают все стилистическое разнообразие документов интернета. Однако классификация, получаемая с помощью дискриминантного анализа, относит любой объект к одному из классов. Поэтому для работы на реальных данных классификацию пришлось дооснастить инструментами для проверки достоверности и принятия окончательного решения о принадлежности документа жанру. Для этого были использованы оценки, основанные на расстоянии Мехаланобиса [2,5]: оценка удаленности объекта от центра класса и апостериорная вероятность принадлежности объекта классу. На основе первой можно сделать заключение о доле объектов класса, которые лежат дальше от центра, чем данный объект. Малые значения говорят о том, что объект сильно отличается от типичных представителей класса. Параметр можно интерпретировать как вероятность отвергнуть правильную классификацию. Второй параметр – апостериорная вероятность принадлежности объекта классу, – вычисляется на основе расстояний объекта до центров всех классов и показывает, насколько объект «близок» именно к этому классу. Параметр можно интерпретировать как вероятность неправильной классификации.

5. Тестирование процедуры классификации

5.1 Тестовая выборка документов

Тестовая выборка документов для проверки полученной процедуры была сформирована следующим образом.

- a) Основой для формирования тестовой выборки стали 2 млн. документов, или 1/30 часть всех документов, проиндексированных Яндексом на начало марта 2002 г. (каждый сервер представлен 1/30 частью своих документов).
- b) Из этих 2 млн. документов выбраны те, что содержат не менее 500 слов (ок. 100 тыс.). Заметим, что доля этих текстов по количеству словоупотреблений составляет 27% от всего множества документов.
- c) Из оставшегося множества выбраны документы, для которых модуль распознавания кодировки Яндекса определил наличие символов в кириллической кодировке – в основном это документы на русском языке и небольшое количество на украинском. После этого осталось около 80 тыс. документов.
- d) Из этого множества случайным образом было выбрано по 100 документов каждого из шести жанров; к этим 600 добавлены 300 документов «без жанра». Жанр или его отсутствие определялось по принципу наследования признаков от вышележащего домена, при этом с одного сервера бралось не более одного документа. В дальнейшем эту классификацию будем называть *базовой*.
- e) Эти 900 документов были «перемешаны», и первые 300 из них были предъявлены пяти редакторам каталога для «ручной» классификации по жанрам.
- f) Для каждого документа жанр определялся следующим образом: если три или более редактора приписали документу один и тот же жанр или его отсутствие, то документу приписывался этот жанр (что произошло в 269 случаях из 300), иначе считалось, что документ жанра не имеет.

г) После этого из выборки было исключено 9 документов, которые состояли в основном из латиницы – в этих случаях модуль распознавания кодировки Яндекса дал неподходящий в данном контексте результат.

Классификацию, построенную таким образом, далее будем называть *эталонной*. Она и будет служить образцом для дальнейшего тестирования. (Попутно было измерено «качество классификации» для каждого из редакторов относительно эталонной классификации – оно колеблется от 79% до 86%.)

5.2 Результаты автоматической классификации

Для сравнения с матрицей классификации обучающей выборки (Таблица 1), в Таблице 2 приведён результат сравнения результата автоматической классификации документов эталонной выборки, имеющих один из известных автоклассификатору четырёх жанров. Всего таких документов 135. Качество по этому подмножеству можно признать удовлетворительным. Заметим, однако, что более всего ухудшилось качество классификации жанра *Советы*.

Таблица 2. Матрица классификации документов четырех базовых жанров

	ХудЛит	НаучТех	НормДок	Советы	Все-авто	Точность
ХудЛит	42	.	.	7	49	0.857
НаучТех	.	31	1	2	34	0.912
НормДок	.	1	19	.	20	0.950
Советы	4	2	2	24	32	0.750
Все-эталон	46	34	22	33	135	0.859
Полнота	0.913	0.912	0.864	0.727	0.859	

Поскольку автоматическая классификация в настоящий момент не различает жанров *НаучПоп* и *НаучТех*, во всех дальнейших тестах эти жанры объединены; далее объединённый жанр называется *Научным*. Результат представлен в Таблице 3. Заметим, что точность определения жанра *НаучТех* от слияния с *НаучПоп* даже увеличилась, хотя полнота уменьшилась. С другой стороны, снова заметно снизилось точность классификации *Советов* из-за ошибочно отнесённых туда 9 документов из *НаучПоп*. Точность *ХудЛит* тоже несколько уменьшилась за счет 4 ошибочно отнесённых к нему документов. Качество классификации *НормДок* от этой операции не изменилось.

Таблица 3. Матрица классификации при объединении двух научных жанров

	ХудЛит	Научный	НормДок	Советы	Все-авто.	Точность
ХудЛит	42	4	.	7	53	0.792
Научный	.	42	1	2	45	0.933
НормДок	.	1	19	.	20	0.950
Советы	4	11	2	24	41	0.585
Все-эталон	46	58	22	33	159	0.799
Полнота	0.913	0.724	0.864	0.727	0.799	0.799

Поскольку процедура автоматической классификации в настоящий момент не отличает жанра *Публицистика*, в дальнейших тестах этот жанр отнесён к выборке *Нет жанра*. В Таблице 4 представлен результат автоматической классификации по полной эталонной выборке в режиме, когда автомат приписывает жанр всем документам. В этом режиме наблюдается неплохая полнота (если исключить из рассмотрения выборку *Нет жанра*) при

низкой точности; падение точности объясняется ошибочным отнесением к жанрам большого количества документов из выборки *Нет жанра*.

Таблица 4. Матрица классификации полной эталонной выборки без оценки достоверности классификации

	ХудЛит	Научный	НормДок	Советы	Нет жанра	Все-авто.	Точность
ХудЛит	42	4	.	7	30	83	0.506
Научный	.	42	1	2	31	76	0.553
НормДок	.	1	19	.	22	42	0.452
Советы	4	11	2	24	49	90	0.267
Нет жанра	0	0.000
Все-эталон	46	58	22	33	132	291	0.436
Полнота	0.913	0.724	0.864	0.727	0.000	0.436	

Для принятия решения о принадлежности документов жанру были использованы параметры оценки удаленности объекта от центра класса и апостериорной вероятности принадлежности объекта классу. Результаты приведены в Таблице 5.

Таблица 5. Матрица классификации полной эталонной выборки с оценкой достоверности классификации

	ХудЛит	Научный	НормДок	Советы	Нет жанра	Все-авто	Точность
ХудЛит	39	1	.	2	13	55	0.709
Научный	.	30	1	1	12	44	0.682
НормДок	.	.	16	.	3	19	0.842
Советы	.	2	2	11	11	26	0.423
Нет жанра	7	25	3	19	93	147	0.633
Все-эталон	46	58	22	33	132	291	0.649
Полнота	0.848	0.517	0.727	0.333	0.705	0.649	

Некоторой побочной целью данного тестирования была оценка эффективности механизма наследования признаков в отношении жанровой классификации. Для такой оценки было проведено тестирование качества базовой классификации, т.е. классификации на основе информации из каталога с применением наследования жанра, относительно эталонной. Результат представлен в Таблице 6.

Таблица 6. Качество базовой классификации относительно эталонной

	ХудЛит	НаучТех	НормДок	Советы	НаучПоп	Публиц	Нет ж.	Все-баз.	Точн.
ХудЛит	26	.	.	.	2	3	2	33	0.788
НаучТех	3	17	1	2	2	3	10	38	0.447
НормДок	.	.	18	.	.	2	3	23	0.783
Советы	1	.	.	21	4	1	7	34	0.618
НаучПоп	4	5	.	3	9	8	12	41	0.220
Публиц	.	1	.	2	.	30	4	37	0.811
Нет ж.	12	11	3	5	7	17	30	85	0.353
Все-эталон	46	34	22	33	24	64	68	291	0.519
Полнота	0.565	0.500	0.818	0.636	0.375	0.469	0.441	0.519	

5.3. Обсуждение результатов тестирования

5.3.1. Качество итоговой классификации. Основное количество ошибок итоговой классификации (см. Таблицу 5) связано с классом *Нет жанра*. Заметим, что этот класс наполовину состоит из документов жанра *Публицистика*. Успешные опыты по автоматическому распознаванию публицистического стиля [1] позволяет надеяться на то, что после выделения *Публицистики* в отдельный жанр эта проблема будет во многом снята.

5.3.2. Качество классификации Советов. Обращает на себя внимание низкое качество классификации *Советов*. Дополнительный анализ ошибок, связанных с *Советами*, показал следующее.

Из 16 случаев потери точности, т.е. ошибочного отнесения к *Советам*, 7 документов – форумы, 2 – *НормДок* на украинском языке, один документ преимущественно в латинице, остальные 6 достаточно разнородны.

Из 22 случаев потери полноты, т.е. ошибочного отнесения *Советов* к другим жанрам, 4 документа – форумы, остальные – в основном советы специалистов в самых разных областях (кормление младенцев, окраска волос, советы горнолыжникам, описания прохождения игр, советы по установке Windows и т.п.).

Можно предложить два варианта модификации жанра *Советы*:

- 1) Считать *Советы* подмножеством *Разговорного* жанра. Тогда дополнением *Советов* в *Разговорном* жанре назвать *Обсуждениями* и попытаться этот класс также выделять автоматически. Недостаток этого варианта – потеря прагматической ценности (трудно будет выделить пособия и рекомендации специалистов в разных областях как отдельную сущность).
- 2) Перенести *Советы* из Жанров в другое фасетное измерение (напр. в *Справочно-информационный* фасет). На его месте основать *Разговорный* жанр.

5.3.3. О Научном жанре. Вероятно, разделение в автоматической классификации *Научного* жанра на *НаучТех* и *НаучПоп* позволит повысить как качество классификации по *Научному* жанру, так и общее качество классификации по жанрам.

5.3.4. Качество базовой классификации. Как видно из таблицы 6, удовлетворительная точность базовой классификации наблюдается в случае жанров *НормДок*, *ХудЛит* и *Публиц.* Это, по-видимому, означает, что документы этих жанров лежат в Сети в виде компактных и однородных совокупностей – *коллекций*, которые относительно легко поддаются «ручной» классификации (примеры: <http://www.kodeks.net>, <http://www.lib.ru> и <http://www.gazeta.ru> соответственно). Удовлетворительная полнота наблюдается только в случае жанра *НормДок*. Это, позволяет предположить, что нормативные документы в Сети представлены в основном в виде коллекций, в то время как остальные жанры часто существуют и в виде «россыпей».

6. Заключение

Фасетная классификация представляется перспективным способом описания ресурсов интернета и хорошей основой организации интернет-каталогов. Фасетная классификация сочетает в себе гибкость, выразительность и простоту.

Значение фасета «Жанр» для отдельного документа и коллекции документов может быть определено на основе автоматической процедуры с приемлемым для практических целей качеством, что демонстрирует разработанная макетная реализация процедуры жанровой классификации документов.

Полученная процедура жанровой классификации обладает гибкостью и широкими возможностями настройки. При необходимости не составляет большого труда перейти к классификации на основе новой структуры классов и нового набора параметров.

В настоящее время продолжается тестирование полученной процедуры жанровой классификации, а также эвристики для определения жанра ресурса на основе жанров составляющих его документов.

7. Благодарности

Авторы выражают благодарность Андрею Целищеву за помощь в программной реализации метода, а также редакторам каталога «Яндекс» Юлии Гребенюк, Евгению Ломизе, Алексею Остроумову, Галине Тихончук и Ирине Угловой за классификацию тестовой выборки документов.

Литература

1. Браславский П.И. Использование стилистических параметров документа при поиске информации в Internet // Доклады VI рабочего совещания по электронным публикациям – EL-PUB-2001. Новосибирск: ИВТ СО РАН, 2001. <http://www.ict.nsc.ru/ws/elpub2001/1812/>
2. Клекка У.Р. Дискриминантный анализ // Факторный, дискриминантный и кластерный анализ: Пер. с англ. М.: Финансы и статистика, 1989. С. 78-138.
3. Прикладная статистика: Классификация и снижение размерности: Справ. изд. / Под. ред. С.А. Айвазяна. М.: Финансы и статистика, 1989.
4. Ранганатан Ш.Р. Классификация двоеточием. Основная классификация. Пер. с англ. / Под. ред. Т.С. Гомолицкой. М.: ГПНТБ СССР, 1970.
5. Рао С.Р. Линейные статистические методы и их применения. М.: Наука, 1968.
6. Спиркин А.Г. Знание // БСЭ, М.: Советская энциклопедия, 1969-1978.
7. Karlgren J. Cutting D. Recognizing Text Genres with Simple Metrics Using Discriminant Analysis // Proc. 15th Int. Conf. on Computational Linguistics (COLING). Kyoto, 1994. Vol. 2. P. 1071-1075.

FACET BASED INTERNET DIRECTORY DESIGN AND AUTOMATED DOCUMENT CLASSIFICATION ACCORDING TO GENRE

P. I. Braslavsky

Key words: Internet directory, faceted classification, automated document classification, genre recognition, stylistic classification, discriminant analysis

The amount and variety of Internet resources lead to complex taxonomies of universal Internet directories. As a possible solution to the problem, a faceted classification (FC) approach can be considered. Yandex Internet directory built using this approach includes the GENRE facet among other facets. The development of an automated procedure for document classification according to genre is described; the results of the procedure testing are cited and discussed. A conclusion about efficiency of the designed procedure is drawn and proposals for further method development and applications are formulated.