нии которых, обладателем таких сведений введен режим коммерческой тайны [3].

Объектом прав использования РИД в составе единой технологии является выраженный в объективной форме результат научно-технической деятельности, который включает в том или ином сочетании изобретения, полезные модели, промышленные образцы, программы для ЭВМ или другие результаты интеллектуальной деятельности, подлежащие правовой охране, и может служить технологической основой определенной практической деятельности в гражданской или военной сфере (единая технология). В состав единой технологии могут входить также результаты интеллектуальной деятельности, в том числе технические данные, другая информация [4]. Однако, необходимо отметить, что такая форма охраны распространяется только на технологии, созданные на бюджетные средства.

Рекомендуем обратить внимание, что один и тот же объект может охраняться с помощью различных способов, иногда взаимоисключающих. Например, техническое решение типа – устройство, может охраняться как объект патентного права (изобретение или полезная модель) или как секрет производства (ноу-хау). Охрана в качестве объекта патентного права требует полного разглашения информации об объекте, тогда как охрана в качестве ноу-хау, напротив, требует сохранения конфиденциальной информации. Некоторые изделия, например оригинальная упаковка, могут быть в равной степени признаны как промышленным образцом, так и объемным товарным знаком.

Выводы

Наиболее эффективной формой охраны объектов интеллектуальной собственности при выведении их на конкурентные рынки является комплексный подход к охране, комбинирующий различные способы охраны объектов. Так, например, процесс или способ создания РИД в области инженерной экологии может охраняться как зарегистрированное изобретение, патентным правом, а описание соответствующей технологии – авторским правом.

Литература

1. Гражданский кодекс РФ. Часть IV, Парламентская газета. 2006, № 214-215, ст. 1259.
2. Гражданский кодекс РФ. Часть IV, Парламентская газета. 2006, № 214-215, ст.1349.
3. Гражданский кодекс РФ. Часть IV, Парламентская газета. 2006, № 214-215, ст. 1465.
4. Гражданский кодекс РФ. Часть IV, Парламентская газета. 2006, № 214-215, ст. 1542.

Анализ текстовой и цифровой информации для моделирования процессов

д.т.н. проф. Софиев А.Э., Верещагин Г.М.
Университет машиностроения
vergleb@yandex.ru

Аннотация. В статье представлено описание простейших классификаторов, а также приведено сравнение популярных алгоритмов категоризации текста с использованием тестовых выборок.

Ключевые слова: классификация информации, цифровой анализ информации

Введение

Ежегодно увеличивается объем существующей в мире информации, и поэтому становится все более актуальной задача автоматического анализа и классификации текстовой информации. Это обусловлено необходимостью иметь возможность поиска по имеющемуся массиву текста. Также это необходимо для того чтобы иметь возможность контролировать перемещение информации по сети между компьютерами.

Виды классификаторов

Для решения этой задачи часто применяются различные тематические классификаторы, рубрикаторы и т.д., которые позволяют производить поиск документов удовлетворяющих некоторым критериями в некоторой информационной базе. Существует несколько видов

классификаторов:

1. «Ручные классификаторы». Классификатор этого типа обычно представляет собой множество рубрик, объединенных в иерархию (рубрикатор). К каждой рубрике приписываются соответствующие ее тематике документы. Иерархия рубрик может являться деревом, однако возможны ситуации, когда некоторые рубрики являются дочерними сразу для нескольких родительских рубрик. Пример: «новости математики» может являться дочерней одновременно для рубрики «математика» и рубрики «новости науки».

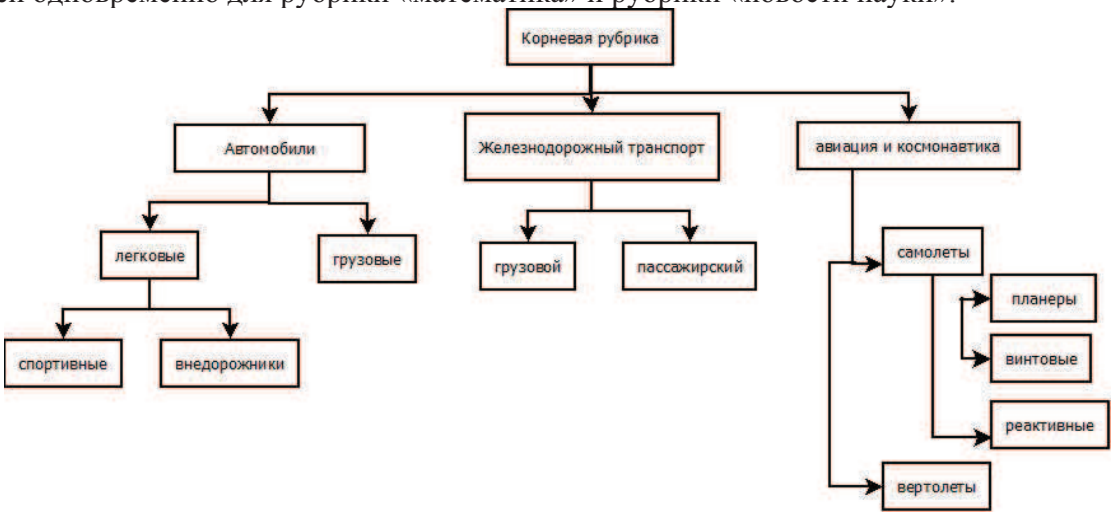

Рисунок 1. Пример рубрикатора

Существенный недостаток классификационного поиска в том, что документы, как правило, приходится классифицировать вручную. То есть при добавлении в массив нового документа сначала нужно его проанализировать и определить, к каким рубрикам классификатора он относится (микропроцессорные системы, сотрудничество компьютерных фирм, изобразительное искусство средневековья и т.д.). После этого документ станет доступным для поиска по классификатору.

Очевидно, что при большом потоке входных документов применение ручной классификации становится очень трудоемким. Обеспечить высокую полноту ручной классификации большого объема документов оказывается сложно даже при помощи большого количества специалистов. Это обусловлено тем, что при ручной классификации часто возникает ситуация, что документ, соответствующий сразу нескольким рубрикам, оказывается приписан только части из них. Обычно количество таких ошибок пропорционально размерности рубрикатора.

2. «Автоматические классификаторы». Классификатор этого типа представляет собой систему, принимающую решение об определении документа в категорию автоматически. Это делается в частности с помощью частотного анализа по заданным ключевым словам.

Существует 3 варианта автоматической классификации:

1. Поиск в искомом тексте лексем из документов обучающей выборки. В данном случае документ А из обучающей выборки и классифицируемый документ Б разделяется на лексемы (словоформы), поиск которых осуществляется в классифицируемом документе. Таким образом поиск дает положительный результат для слов в разных падежах, а также однокоренных слов, присутствующих в обоих документах. Доля найденных лексем обучающего документа среди всех лексем принимается за вероятность принадлежности документа к категории.

2. Полнотекстовый поиск. В этом случае осуществляется поиск слов из исходного документа в классифицируемом без учета лексем(словоформ). Такой поиск даст положительный результат для слова только в случае полного совпадения. Все слова из

обучающего документа А сравниваются со всеми словами из классифицируемого документа Б. Доля общих слов документов А и Б есть вероятность принадлежности документа к категории.

3. Поиск по строгому соответствию. Это самый простой и малоэффективный тип поиска. В нем положительный результат возможен только если есть полное соответствие фрагментов текста документов А и Б.

Сравнение алгоритмов

Сравним эффективность трех вышеуказанных алгоритмов с помощью ЭВМ на примере трех наборов обучающих документов:

1. Коллекция новостных сообщений с сайтов RBK и Инфоарт.

2. Фрагмент юридической базы Консультант-плюс.

3. Аннотации к ресурсам, участвующим в рейтинге top100 Рамблера. Отличительной особенностью данного набора является наличие в нем поискового спама (специальных текстов, предназначенных для повышения позиции сайта в выдаче поисковой машины Рамблер).

Характеристики наборов показаны в таблице 1.

Таблица 1

Характеристики наборов данных для анализа

№ базы	рубрик	документов	Примечание
1	243	14403 (140 Мб)	новости РБК, ИнфоАрт
2	29	1590 (155 Мб)	часть базы Консультант+
3	57	101013 (42 Мб)	каталог Rambler s top100

Для оценки качества классификации будем использовать другой набор документов (тестовый), содержащий соответственно случайно выбранные новостные статьи, юридические документы, и описания к сайтам, не содержащиеся в первом наборе документов.

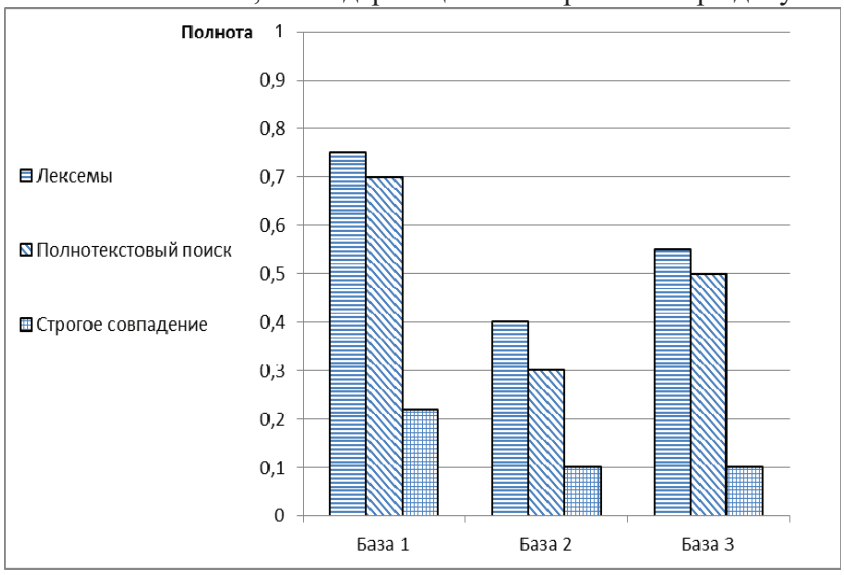

Рисунок 2. Зависимость полноты классификации от способа сопоставления терминов

Проведем следующий эксперимент: попарно (новости с новостями и.т.д.) сравним тестовый набор документов с исходной базой, среднюю вероятность принадлежности документа к категории будем использовать в качестве оценки полноты алгоритма. За полноту принимается доля от общего числа соответствующих документов которые распознались классификатором.

Также проанализируем точность классификации. Точность зависит от процента ошибок второго рода, то есть когда документ был ложно приписан к категории. Чем выше точность, тем меньше таких ошибок.

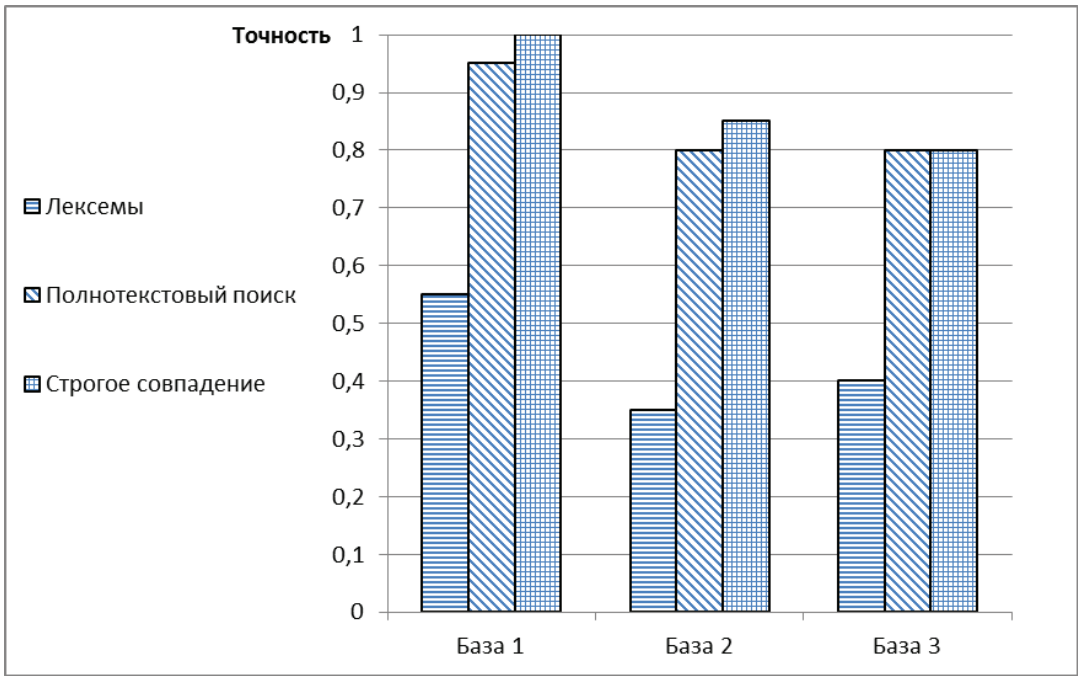

Рисунок 3. Зависимость точности классификации от способа сопоставления терминов

Выводы

Как видно из диаграмм, классификатор, использующий лексемы дает большую полноту при ощутимо меньшей точности. Это в основном связано с многозначностью терминов, попавших в семантически образы рубрик. Классификация с использованием полнотекстового поиска эффективнее, процент ошибок меньше. Использование поиска по строгому совпадению дает максимальную точность, но количество ошибок первого рода очень велико.

Различия между методами максимальны для больших тестовых наборов, как видно на примере набора №3, различия для которого минимальны в силу его малого размера.

Эффективность классификации в зависимости от метода сопоставления показана в следующей таблице:

Таблица 2

Зависимость эффективности классификации от метода сопоставления.

Метод	набор № I	набор №2	набор №3
Лексемы	0,62	0,36	0,45
Полнотекстовый поиск	0,81	0,45	0,61
Строгое совпадение	0,36	0,16	0,16

Таким образом эффективность классификации максимальна при полнотекстовом поиске словосочетаний, который обеспечивает баланс между полнотой, характерной для традиционных методов, использующих однословные термины, и точностью, характерной для традиционных методов, использующих многословные термины.

Литература

1. Korde V. Text Classification and Classifiers:A Survey //International Journal of Artificial Intelligence & Applications (IJAIA), Vol.3, No.2, March 2012.
2. B.S. Harish, S. Manjunath, D.S. Guru "Text document classification: An approach based on indexing. International Journal of Data Mining & Knowledge Management Process (IJDKP) Vol.2, No.1, January 2012.
3. Добрынин В.Ю., Клюев В.В. Некрестьянов И.С. Оценка тематического подобия текстовых документов // Электронные библиотеки: перспективные методы и технологии: Вторая всероссийская научная конференция. - Санкт-Петербург, 2000.

К расчёту процесса очистки промышленных газов от диоксида углерода

Гуреев А.О., к.т.н. доц. Пикулин Ю.Г.
Университет машиностроения
gureev_aleksei@mail.ru

Аннотация. Предложена математической модель абсорбера для малых объёмов очищаемых газов. Проведён расчёт основных размеров для опытно-промышленной установки, выполнен расчёт основных размеров абсорбера для промышленной установки, который показал работоспособность составленной математической модели для многотоннажного производства.

Ключевые слова: очистка газов, абсорбер, моделирование.

В крупнотоннажных производствах очистка газов от кислых компонентов производится как в технологической цепочки получения того или иного продукта, так и при выбросе дымовых газов какого-либо производства. Количество извлекаемых компонентов из газовых смесей исчисляется сотнями тысяч тонн в год. Одним из самых дешёвых способов очистки является циркуляционный способ, в котором после насыщения абсорбент направляют на стадию регенерации, где, в том числе, происходит выделение поглощённого компонента в чистом виде. Количество циркулирующего абсорбента в системе зависит от его абсорбционной (поглотительной) ёмкости.

Вследствие разнообразия топок, котельных и других аналогичных устройств, очень широка сфера приложения различных методов контроля чистоты выбросов. Специалист в данной области имеет возможность выбрать оптимальный вариант или найти способы улучшения уже функционирующих конструкций. Общий интерес представляет применимость отдельных методов к конкретным типам загрязняющих выбросов, их универсальность, экономичность, перспектива усовершенствования, увеличение производительности и возможные недостатки. Регенерация абсорбента проводится, как правило, либо путём сброса давления – дросселированием, либо при нагревании.

Существуют разнообразные методы очистки газа от CO_2 и совместной очистки от CO_2 и других примесей [1]:

1. Водная очистка – наиболее простой и известный метод удаления CO_2. Недостатком данного метода является большой расход электроэнергии, т.к., как правило, абсорбция проводится при высоком давлении, а регенерация – при его сбросе.
2. Очистка газов водными растворами полиаминов. Недостатком является их высокая коррозионная активность.
3. Очистка газов раствором дигликольамина (ДГА). Недостатком является высокая стоимость растворителя и его сравнительно большие потери.
4. Очистка растворами карбонатов.

А) Очистка горячими растворами поташа. Основной недостаток – сильная коррозия оборудования. При сравнении с процессом МЭА–очистки концентрация CO_2 в очищенном га-