А.С. Мохов, В.О. Толчеев, В.Е. Туманов

Классификация научных публикаций в области химической физики по русскоязычным и англоязычным названиям

Рекомендуемая форма библиографической ссылки

Мохов А.С., Толчеев В.О., Туманов В.Е. Классификация научных публикаций в области химической физики по русскоязычным и англоязычным названиям // Научный сервис в сети Интернет: труды XVIII Всероссийской научной конференции (19-24 сентября 2016 г., г. Новороссийск). — М.: ИПМ им. М.В.Келдыша, 2016. — С. 279-283. — doi:10.20948/abrau-2016-19

Размещена также презентация к докладу

Классификация научных публикаций в области химической физики по русскоязычным и англоязычным названиям

А.С.Мохов[1], В.О.Толчеев[1], В.Е.Туманов[2]

1 НИУ «Московский энергетический институт»
2 Институт проблем химической физики РАН

Аннотация. В статье рассматривается задача анализа и обработки базы данных ИПХФ РАН, в которой хранятся названия научных статей и отчетов по НИР, представленные на русском и английском языке. С помощью специального программного комплекса TextCat решаются задачи выявления и удаления полных дубликатов, а также обучения классификаторов по заданным тематикам для дальнейшего упорядочивания части публикаций из БД и организации их быстрого поиска и просмотра.

Ключевые слова: Интеллектуальный анализ текстовых данных, информационный поиск, классификация двуязычных библиографических документов, цифровые библиотеки.

1. Введение

В области обработки и анализа научной информации (статьи в журналах, доклады на конференциях, отчеты по НИР, диссертации, монографии) используется два основных способа представления публикаций: в виде полнотекстового документа и в виде библиографического документа.

Полнотекстовый документ обычно состоит из названия, фамилии и имени автора, места его работы, краткого описания (аннотации), ключевых слов, основного текста, который делится на главы (абзацы), а также ссылок на литературу. *Библиографическое описание* (БО) представляет собой часть полнотекстового документа, опубликованного в реферативном журнале или сохраненного в специализированной базе данных. БО включает название (*title – T*), аннотацию (*abstract – A*), ключевые слова (*keywords – K*) и некоторые другие поля (фамилия и имя автора, место работы и т.п.), которые достаточно редко используются при классификации текстов.

Различия между полнотекстовыми и библиографическими документами обсуждаются по разным аспектам. В частности, отмечается, что библиографическое описание содержит значительно меньшее количество терминов и является более формализованным, что значительно облегчает разработку процедур для обработки и анализа данных [1,2]. Более того, работа с

БО не только экономит время на просмотр публикаций и понимание их смысла, но и коммерчески более выгодна, так как значительная часть библиографических документов распространяется в сети Интернет бесплатно. Проанализировав содержание БО, исследователь может более прицельно заказать те платные полнотекстовые версии, которые в наибольшей степени отвечают его научным интересам.

Наряду с обработкой и анализом полнотекстовых и библиографических документов на практике встречаются и другие узкоспециализированные задачи, к числу которых относится классификация научных публикаций по названиям. Такая задача возникает, в частности, при анализе больших ретроспективных баз данных (БД), накопленных научными организациями за время своей деятельности. Эти БД редко преобразуются в полнотекстовые (или библиографические), так как подобные трансформации являются крайне трудозатратными и ресурсоемкими, требуют оцифровки больших объемов текстовых данных. В связи с этим возникает необходимость обработки чрезвычайно коротких документов, в частности, их упорядочивание по тематикам, устранение дубликатов, извлечение сведений для наукометрического анализа.

Для решения задачи обработки чрезвычайно коротких документов необходимо особенно тщательно выбирать программно-алгоритмический инструментарий, способный обеспечивать высокую точность в условиях ограниченного размера словаря (количество доступных терминов из названий на порядок меньше, чем количество терминов той же выборки, содержащей библиографические описания). Кроме того, используемые алгоритмы должны быть инвариантными к языку публикаций, так как названия статей в БД могут приводится как на английском, так и на русском языках.

В данной статье для обработки и анализа научных статей, представленных своими названиями, применяется программный комплекс (ПК) *TextCat*, разработанный на кафедре Управления и информатики НИУ «Московский энергетический институт». ПК *TextCat* предназначен для обработки двуязычной (русско-английской) научной информации в интересах специалистов-предметников (и небольших коллективов пользователей – отделов, лабораторий, кафедр). В ПК реализована возможность выявления информативных терминов каждого класса на основе расчета совстречаемости термина и класса в документах выборки. Это позволяет формировать профили классов (терминологические портреты, семантические образы) из наиболее специфичных, классообразующих терминов.

2. Анализ базы данных ИПХФ РАН с помощью программного комплекса *TextCat*

В данной статье для анализа используется база данных Института проблем химической физики РАН (ИПХФ РАН), в ней содержатся русскоязычные и англоязычные названия статей, которые сотрудники

института опубликовали в российских и зарубежных журналах. Названия статей отражают основные направления научных исследований ИПХФ РАН, многие из них соответствуют тематикам работ широко известных научных школ, сформировавшихся в институте.

В ходе проведения исследований решались следующие задачи:

- Обучение классификаторов по указанным тематикам (классам). Применение обученных классификаторов для отнесения тематически близких документов, находящихся в БД, к выделенным классам, т.е. упорядочивание части публикаций из БД и организация их быстрого просмотра и анализа.

- Выявление и удаление полных дубликатов, которые появляются из-за ошибок при вводе информации в БД, а также из-за наличия совпадающих названий, представленных в БД на разных (русском и английском) языках. Для выявления таких дубликатов требуется привести все названия в БД к единому языку (например, с помощью машинного перевода) и проанализировать их на наличие полностью идентичных последовательностей слов.

- Проверка работоспособности ПК *TextCat* и оценка точности реализованных методов в условиях обработки и классификации чрезвычайно коротких текстовых документов, а также обоснование применимости ПК *TextCat* для анализа научных публикаций в области химической физики.

Нами для проведения исследований и решения вышеуказанных задач выбраны девять крупных направлений (классов), по которым в ИПХФ РАН активно проводятся научные работы:

- Структурная химия и кристаллохимия.
- Полимеры и композиционные материалы.
- Каталитические превращения олефинов.
- Экстремальные состояния вещества.
- Термодинамика высокотемпературных процессов.
- Металлополимеры
- Спектроскопия и наноматериалы
- Химия углеводородов
- Ионика твердого тела

Размер обучающего и экзаменационного множества составляет 1035 и 315 (по 115 обучающих и 35 экзаменационных документов в каждом классе). Из различных сочетаний указанных тематик было сформировано по три обучающие и экзаменационные выборки одинакового размера по 7 классов в каждой, и по ним рассчитывались ошибки (и определялась средняя ошибка). В выборки включены названия статей, как на русском, так и на английском языке (примеры документов из обучающей выборки приведены на рис.1).

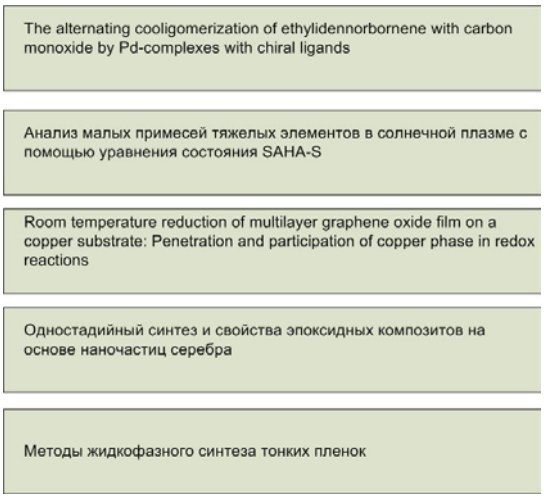

Рис.1. Примеры документов.

При формировании выборок и определении метки класса использовались эксперты – сотрудники ИПХФ РАН.

С помощью ПК *TextCat* проводилась предварительная обработка двуязычных (русских и английских) публикаций, которая включала удаление стоп-слов и стемминг.

В связи с крайне малым размером текстов введено ограничение на число слов в отбираемых документах. Тексты, которые содержали менее 3 слов после проведения стемминга и отсечения стоп-слов, из выборок исключались (во всех выборках было обнаружено три такие статьи).

На основе использования реализованных в ПК *TextCat* базовых профильных методов (РО-профиль на основе статистического χ^2-критерия, НМИ-профиль на основе нормированного критерия взаимной информации и J-профиль (коэффициент Жаккара)), а также модификаций базовых профильных методов (UNI5 и UNI6) проводилось построение профилей классов [1,2].

Результаты классификации на экзаменационных выборках показали, что наилучшая точность показана методом UNI6 (см. таблицу 1).

Таблица 1. Средние, минимальные и максимальные ошибки классификации

Метод	***Средняя ошибка, %***	***Мин. ошибка, %***	***Макс. ошибка, %***
РО	20,27	19,18	22,04
НМИ	20,68	18,36	24,08
J	23,13	21,64	24,89
UNI5	21,22	20,40	22,44
UNI6	19,86	18,36	21,22

Отметим, что ошибка классификации существенно снижается при использовании вместо названий статей их библиографических описаний (название, аннотация, ключевые слова). Так, при использовании БО средняя

ошибка методов, представленных в таблице 1, находится в интервале от 12,37 до 15,08. Кроме того, для увеличения точности классификации по названиям целесообразно проводить их перевод на второй язык и осуществлять классификацию полученного более информативного (русско-английского) текста. Это позволяет уменьшить ошибку классификации на 2-3 процента.

Применительно к БД ИПХФ РАН решалась еще одна задача, которая заключалась в поиске и удалении полных дубликатов. Наибольшие проблемы возникали из-за наличия в БД русскоязычных (или англоязычных) названий, которые полностью идентичны в случае их представления на одном языке. Для выявления таких совпадающих названий нами применялся машинный перевод и приведение анализируемых названий к единому (русскому) языку, для исключения ошибочного решения из-за погрешностей перевода использовались экспертные оценки специалистов-предметников. Необходимо отметить, что задача ограничивалась именно поиском полных дубликатов, так как поиск и выявление нечетких дубликатов (почти дубликатов) не целесообразны в условиях анализа чрезвычайно коротких документов, содержащих только названия статей. Для выявления дубликатов применялся коэффициент Жаккара, который также используется в ПК *TextCat* в качестве одного из способов выявления информативных терминов и составления профилей классов. Вместе с тем коэффициент Жаккара хорошо себя зарекомендовал в качестве эффективного средства выявления (нечетких) дубликатов [1] и широко используется на практике. Выявленные дубликаты специальным образом помечаются в БД, что дает дополнительные возможности при проведении наукометрических исследований, например, исключать из рассмотрения статьи с одинаковыми названиями на двух языках.

3. Заключение

Проведенные исследования показали эффективность применения ПК *TextCat* для обработки коротких научных документов, заданных своими названиями, в области химической физики. Представляется целесообразным расширить число классов, по которым проводится упорядочивание документов в БД, с учетом структуры ИПХФ и информационных потребностей отделов и лабораторий.

Литература

1. Мохов А.С., Толчеев В.О. «Способы учета структуры научных документов в задачах обработки и анализа текстовой информации» // Информационные технологии, № 5, 2016, с. 332-339.
2. Мохов А.С., Толчеев В.О. Разработка профильных методов классификации двуязычных текстовых документов // Материалы 6-й Всероссийской мультиконференции по проблемам управления. ИПУ РАН и ЮФУ. Дивноморское. 2013, Том 1, с. 75-79.