УДК 025.4.026
DOI 10.33186/978-5-85638-221-0-2019-46-48

Некоторые аспекты оптимизации обработки документов из электронных ресурсов на иностранных языках при подготовке тематически структурированного потока научных документов

Processing documents from the digital resources in foreign languages for preparing topically structured flow of scientific documents: Several optimization aspects

Н. А. Чуйкова
Всероссийский институт научной и технической информации
Российской академии наук,
Москва, Россия

Nadezhda Chuikova
All-Russia Institute for Scientific and Technical Information,
Russian Academy of Sciences,
Moscow, Russia

Рассмотрены возможности оптимизации обработки документов из электронных ресурсов на иностранных языках при подготовке тематически структурированного потока научных документов.
В исследования использованы статистические (и, в частности, инфометрические) методы оценки использования первичных научных документов и сериальных изданий (СИ) при производстве вторичных информационных продуктов.
Исследован массив 292 наименования СИ на английском языке, получаемых в электронном виде.
Выявлены две группы СИ: технологически неэффективные и наиболее продуктивные.
Продуктивность оценивалась как степень использования научных документов и СИ при подготовке тематически структурированных массивов научных документов.
На основе оценки тематических распределений и статистических оценок была выделена групп СИ наиболее пригодных для автоматической рубрикации текстов научных электронных документов на английском языке.

The ways to improve processing of documents from the digital resources in foreign languages when preparing topically structured flow of scientific documents are examined.
The study is based on statistical (in particular, infometrical) methods for assessing primary scientific documents and serials when generating secondary information products. The array of 292 SI titles in English available digitally is explored. Two SI groups are identified: technologically non-efficient and most effective. The effectiveness was defined as the intensiveness of use of scientific documents and SI in preparing topically structured arrays of scientific documents. Through evaluating topical distribution and statistical assessments, the IS group most suitable for computer-aided rubrication of texts in English is specified.

Основные цели обработки входного потока научной литературы (НТЛ)

1. Получение тематически структурированного по 18 научным направлениям массива отдельных документов (статья, патент, автореферат и т.п.), подходящих, как по тематической направленности, так и по качественным характеристикам, научных публикаций.

2. Двухуровневая экспертиза. Первичная необходима на этапе выявления тематики документа, тематическое соответствие профилю информационной продукции. Вторичная экспертиза происходит на этапе отбор документов для реферирования.

Экспертиза отвечает за качественные характеристики документов в соответствии с критериями: не использовать вообще; поместить только в электронный каталог; передать на реферирование; отразить во вторичных информационных продуктах ВИНИТИ РАН (РЖ и БД).

3. Ведение Рубрикаторов и Классификаторов.

Основные задачи технологии

- Регистрировать, каталогизировать НТЛ.
- Управлять, контролировать (БО, дублетность, маршруты движения, участники обработки, параметры выработки сотрудников и др.).

- Разделять и фильтровать потоки НТЛ.
- Определять «качество» первоисточников.
- Тематически «экспертировать» и структурировать входной поток НТЛ.
- Создавать и поддерживать автоматизированные рабочие места (регистратора, библиографа, эксперта, референта, редактора и др.)

Области технологии, требующие оптимизации. Причины

Одна из причин – необходимость обработки неравномерных потоков документов в условиях временных, материальных, финансовых и кадровых ограничений.

Перечислим наиболее трудоемкие интеллектуальные этапы подготовки тематически структурированной вторичной информации: тематическая экспертиза, научное реферирование, индексирование, редактирование.

Технологические этапы

1. Тематическая экспертиза

Тематическая экспертиза первого уровня применяется ко всем поступающим документам. Здесь наиболее трудоемким является обработка документов на иностранных языках по наиболее современным проблемам (с новыми терминами и словосочетаниями). Такие тексты часто вызывают затруднения тематической экспертизы т.к. для иноязычных документов из электронных ресурсов для принятия экспертного решения представлены только метаданные. К тому же, зачастую, такой реферат составлен не носителем языка и, во многих случаях, с использованием современных автоматизированных средств перевода, что затрудняет понимание специалистами тематики такого документа.

Общие статистические параметры входного потока НТЛ за 2018 г.

Сериальные издания (СИ) – всего 6,44 тыс. наименований; ИКТ – чуть более 10 тыс. наименований. Всего поступило и зарегистрировано более 53 тыс. основных экземпляров научно-технической литературы (исключены дубли и патенты). Обработано за год более 800 тыс. документов (статей). А также 3292 патентов (из них: российских 8395, США 4897).

Статистические параметры входного потока НТЛ за 2018 г. по документам на иностранных языках.

По сводным статистическим данным всего в 2018 г. из стран дальнего зарубежья поступило более 3 тыс. наименований (СИ и ИКТ), включающих 26,3 тыс. экземпляров.

Из них 2,5 тыс. СИ получены из электронных изданий «Удаленного доступа».

В основе оптимизационных решений лежит применение статистических методов исследования потока документов из сериальных изданий (СИ) для технологической обработки на этапе тематической экспертизы.

Целевыми объектами статистики являются:

1. Распределения СИ по тематикам. Распределение строится на основании тематического признака отдельного документа, далее от документа к экземпляру и к СИ. Такой поход к тематической структуре СИ уникален в своей реализации и ценен.
2. Выявление количественных связей наименований СИ с выпусками РЖ (ядра, рассеяние и т.д.).
3. Выявление изданий, в большей степени пригодных для автоматического классифицирования без участия человека. Степень пригодности определяется как технологический параметр.

Количественные характеристики потока СИ позволяют управлять затратами на производство. При этом «уровень продуктивности» может выступать технологическим параметром.

4. Выявление параметров неэффективности.

Дробление на несколько публикаций (в научной среде для определения этого процесса появился термин salami slices) – аналогичные исследования применяются к различным аналогичным друг другу объектам и т.п. Тем самым, с точки зрения научного информирования, такие статьи следует объединять и обобщать.

При дроблении публикаций в первоисточниках возрастает рассеяние информации, в то время как производство вторичных информационных продуктов должно быть направлено на концентрирование, структурирование и сжатие информации

5. Выявление непродуктивных СИ.

Непродуктивные СИ экономически обрабатывать невыгодно. К ним можно применить автоматизированное классифицирование \рубрицирование документов с дальнейшим помещением их в электронный каталог (с указанием метода определения тематики и процентной степени соответствия тематики на основании примененных алгоритмов).

На базе ВИНИТИ РАН разработаны алгоритмы, «обученные» на массиве информационных данных прошлых лет, определять тематику документов с той или иной степенью точности.

Соответствующие программные интерфейсы внедрены в технологические процессы в т.ч. , в качестве подсказки на этапе тематической экспертизы.

Промежуточные результаты

Исследован массив 292 наименования СИ на английском языке, получаемых в электронном виде.

Выявлены две группы СИ: технологически неэффективные и наиболее продуктивные.

Продуктивность оценивалась как степень использования научных документов и СИ при подготовке тематически структурированных массивов научных документов.

На основе оценки тематических распределений и статистических оценок была выделена групп СИ наиболее пригодных для автоматической рубрикации текстов научных электронных документов на английском языке.