УДК 004.4’242

Открытая система обработки текстов в наукометрии и библиометрии

Л.С. Ломакина,
А.С. Суркова,
Д.В. Жевнерчук,
О.С. Рассадин
Нижегородский государственный технический университет им. Р.Е. Алексеева,
Нижний Новгород, Россия

Аннотация:
В работе рассматривается модель открытой информационной системы анализа и обработки текстов в приложении к задачам наукометрии и библиометрии, а также особенности ее реализации на основе сервис-ориентированной платформы быстрого прототипирования Java Spring MVC «VerliOKa». Особенностями предлагаемого решения являются возможность реализации базовых алгоритмов анализа и обработки текстовых данных в виде RESTful сервисов, выполнения их в распределенной среде, формирование на их основе комбинированных алгоритмов, а также возможность построения интерфейса прикладного программирования для создания клиент-серверных систем с мобильными и стационарными клиентскими подсистемами в сфере науки и образования.

Ключевые слова: библиометрия, сервис-ориентированная система; анализ и обработка текста; RESTful сервисы.

Open system of text processing for scientometric and bibliometric purposes

L.S. Lomakina,
A.S. Surkova,
D.V. Zhevnerchuk,
O.S. Rassadin
Nizhny Novgorod State Technical University named after R.E. Alekseeva,
Nizhny Novgorod, Russia

Absract:
The paper examines the model of an open information system for text analyzing and processing for solving scientometric and bibliometric problems, as well as special aspects of its implementation on the basis of the service-oriented platform for rapid prototyping “Java Spring MVC VerliOKa”. The key features of the proposed solution are the ability to implement basic algorithms for text data analyzing and processing in the form of RESTful services, executing of these processes in a distributed processing environment, creating combined algorithms based on them, and also the ability to build an applied programming interface for creating client-server systems with mobile and fixed client subsystems in the field of science and education.

Keywords: bibliometry, service-oriented system; text analysis and processing; RESTful services.

DOI:10.31432/1994-2443-2018-13-1-34-38

Наукометрия и библиометрия — относительно молодые направления исследований, определяющие совокупность количественных (математических, статистических) методов изучения науки и книг в целом как информационных процессов [1]. Статьи, монографии, тезисы являются источниками слабоструктурированных данных, и одним из подходов по их упорядочиванию, выявлению информационных связей и обеспечению эффективного поиска является индексирование, предполагающее создание и развертывание в единых информационных пространствах, в глобальных сетях на основе информационно-телекоммуникационных систем. При этом сами источники информации хранятся в файловых хранилищах в стандартизированных форматах, а сведения о содержании и статистические показатели, как правило, располагаются в базах данных. Тем не менее, в процессе индексирования структурируется только часть полезной информации, что обуславливает важность, значимость и необходимость развития методов анализа и обработки слабоструктурированных текстовых данных.

Текстовый способ представления информации в информационно-телекоммуникационных системах играет важнейшую роль как на системном, так и на прикладном уровнях.

Как уже отмечалось неструктурированный текст является основным источником статистических и семантических характеристик, используемых в наукометрии, причем известные модели хранения и упорядочивания документов [2,3], содержащих текстовые источники, позволяют организовать к ним стандартизированный доступ практически из любой точки глобальных сетей с персонального или

мобильного вычислительного устройства. Такая доступность публикаций с одной стороны расширяет кругозор исследователей, формирует единую исследовательскую среду в масштабах мирового научного сообщества, но с другой — стимулирует рост плагиата. Поэтому актуальными являются задачи: идентификации авторов (индивидуальных и коллективов) научных текстов и распознавания псевдонаучных трудов.

Еще одной актуальной задачей в наукометрии и библиометрии, требующей развития методов аналитической обработки неструктурированных текстов, является распознавание значимых фактов, событий, в научном мире, отражаемых в социальных сетях и тематических порталах новостных лент, информационных системах и сообщениях, которые могут быть использованы для прогнозирования направлений развития науки в той или иной сфере, а также для построения информационных рекомендательных систем.

В рамках настоящей работы предлагается создание открытой информационной системы анализа и обработки текстов (ИСАТ), позволяющей реализовать комплекс методов решения задач анализа текстов разных типов (художественных, научных, интернет текстов, текстов программ) и обладающей следующими возможностями:

- выбор моделей представления текстов;
- выбор методов и алгоритмов обработки текстов;
- формирование критериев, событий, условий применения каждого из методов и алгоритмов;
- отбор существующих программных реализаций в виде встраиваемых библиотек или сетевых сервисов;
- определение системных требований и технологий использования в качестве компонент интегрированных решений;
- определение последовательности выполнения Интернет-сервисов.

Из описания возможностей следуют требования к обработчикам текстовых данных, реализующих простые методы:

а) каждый обработчик может быть применен индивидуально, либо в комплексной модели с другими обработчиками,

б) разные обработчики могут быть реализованы разными группами разработчиков, в) обработчики ориентированы на обработку текстовых данных, которые могут быть представлены в различных форматах, хранятся в узлах глобальных и локальных вычислительных сетей, построенных на основе различных программно-аппаратных платформ.

В связи с этим особую значимость приобретают интеллектуальные системы, которые включают сервисы анализа и обработки текстов. Существующие ресурсы — тезаурусы (WordNet [4], SMART [5, 6]), словари [7], инструменты [8] решают основные задачи обработки и анализа текстов, однако они не могут быть адаптированы к типам, тематикам и другим характеристикам рассматриваемых текстов. Еще одним недостатком большинства существующих сервисов обработки текстов является их ориентация на английский язык.

В основе предлагаемого подхода лежит идея комбинирования базовых алгоритмов, реализованных в форме интероперабельных RESTful-сервисов. Естественным образом вытекает, что ИСАТ должны быть разработаны на основе принципов построения открытых сервис-ориентированных систем [9,10].

Сформулированы технические и функциональные требования к ИСАТ:

1. ИСАТ представляет собой масштабируемый, легко расширяемый комплекс web-сервисов, доступных внешним системам посредством REST API.
2. Клиент серверное взаимодействие осуществляется по защищенному каналу на основе HTTPS, предоставление доступа к ресурсам осуществляется по протоколу OAuth.
3. ИСАТ поддерживает взаимодействие с реляционными СУБД посредством слоя объектно-реляционного отображения.
4. ИСАТ поддерживает базу знаний методов обработки текстовых данных, реализованную на основе технологий Semantic Web, в том числе поддерживаются OWL, RDF.
5. ИСАТ обладает возможностью настройки работы с online словарями, выполненными как web-сервисы [7], а также обеспечивает хранение словарей в своих структурах данных.
6. ИСАТ обладает возможностью поиска и классификации Интернет-текстов по типу, по размеру, по формату хранения и другим свойствам.
7. ИСАТ осуществляет многоуровневую кластеризацию потоковых текстовых данных, предполагающую возможность добавления условия нечеткости и постоянно продолжающегося обучения.
8. ИСАТ обладает генераторами сервисов многофакторного анализа разнотипных

текстовых данных на основе базы знаний простых методов обработки текста.

Функционирование Интернет-сервиса разделено по следующим уровням:

1. Уровень контроллера. Контроллер отвечает за взаимодействие пользователя с одностраничным приложением сервиса. На этом уровне реализована привязка метода для получения текстовых данных к элементу одностраничного приложения. На вход метод принимает текстовый объект, выполняет обращение к методам уровня модели для предварительной обработки текста, подготовки данных для ввода в PHC — отображение слов на индексы, построения модели текста на основе PHC и к уровню сервисов для добавления слов в БД.
2. Сервисы. Данный уровень связан с уровнем Data Access Object (DAO) и уровнем модели. Отвечает за взаимодействие с базой данных через уровень DAO, а именно — вызовы методов для получения текста, списка слов, а также для записи их в базу данных.
3. DAO регулирует отображение объектов на данные БД путем имплементации интерфейса взаимодействия с БД.
4. Модели. Представляют собой классы сущностей EntityText, EntityWord, соответствующие одноименным таблицам базы данных.

На основе предложенного подхода был разработан прототип Интернет-сервиса, предназначенного для построения моделей текста с использованием рекуррентной нейронной сети (PHC) [11]. PHC успешно применяется при решении задач обработки естественного языка.

Обработка сервисом текстовых данных разделена на этапы:

- ввод текста,
- предварительная обработка текста,
- подготовка текстовых данных для передачи в модель,
- построение модели текста,
- добавление новых слов в словарь,
- вывод полученных параметров модели.

На примере описанной модели PHC был создан RESTful сервис определения вероятности появления следующего слова в предложении. Прототип сервис-ориентированной системы анализа и обработки текста реализован на основе платформы быстрого прототипирования Java Spring MVC «VerliOKa», разрабатываемой на кафедре «Вычислительные системы и технологии» НГТУ им. Р.Е. Алексеева.

Было проведено тестирование возможности встраивания сервиса в комплексные системы анализа и обработки текстов, а также его интероперабельность с клиентскими системами: мобильные устройства, ноутбуки и нетбуки, стационарные персональные компьютеры.

Тестирование сервиса проводилось по следующим направлениям: поддержка клиентских платформ, форматы и формы представления текста, режимы доступа, возможность комбинирования компонент (табл. 1).

Таблица 1

Экспериментальное тестирование платформы и RESTful сервиса

1	Поддержка клиентских платформ	Смартфоны, планшеты, ноутбуки, нетбуки, персональные компьютеры	Браузеры: Google Chrome, Mozilla FireFox, IE 9+, Opera
2	Форматы и формы представления текста	Исходный текст: HTML5, PDF Модельное представление: XML, json	
3	Режимы доступа к тексту	По ссылке на удаленный ресурс (HTML 5, PDF), из локального файла, из таблицы базы данных	

Для тестирования сервиса было разработано одностраничное клиентское приложение, которое на уровне платформы является ресурсом, загружаемым по запросу в Web-браузер. Приложение предоставляет механизмы ввода исследуемого текста и вывода на экран его модельных параметров, таких как: весовой коэффициент обратной связи, вектор весов коэффициентов выхода сети, вектор весовых коэффициентов входа сети. Текст может быть представлен в форматах HTML, PDF и предоставлен по прямой ссылке на удаленное файловое хранилище, либо в качестве источника текстовых данных может выступать реляционная база данных. Во втором случае текст извлекается с использованием SQL-запросов.

Экспериментальное исследование сервиса показало:

1. Построенный сервис корректно взаимодействует с такими контейнерами клиентских приложений как браузеры Google Chrome, Mozilla FireFox, IE 9+, Opera.

2. Текст может быть представлен в форматах HTML5 и PDF. Для извлечения полезных текстовых фрагментов из структуры HTML применимы тэг-фильтры, а также информационно-поисковые блоки, анализирующие HTML данные как объектную модель документа (DOM). При работе с текстами в PDF формате, они должны быть представлены в версии от 1.2, это позволит применять компоненты, также ориентированные на работу с объектной моделью PDF документа, например, pdfBox[12].

3. Комплексный сервис может быть собран из компонент, расположенных на удаленных узлах локальных и глобальных вычислительных сетей. Взаимодействие осуществляется по протоколу HTTP через систему интерфейсов. Входные/выходные данные передаются в формате json.

4. Комплексный сервис может быть собран из компонент, расположенных на одном узле, взаимодействие происходит через механизм делегирования. Входные/выходные данные передаются в формате json.

Выводы

В работе была сформулированы технические и функциональные требования к сервис-ориентированной системе анализа и обработки текстов, построен ее опытный образец на основе сервис-ориентированной платформы быстрого прототипирования Java Spring MVC «VerliOKa». Проведены экспериментальные исследования Интернет-сервиса, предназначенного для построения моделей текста с использованием рекуррентной нейронной сети (РНС), в частности, его интероперабельность в открытой клиент-серверной среде, расширяемость по форматам представления текстовых данных, возможности построения комплексных систем анализа и обработки текстов на основе RESTful сервисов.

Таким образом, предложенное решение является открытой информационной системой с многоуровневой архитектурой, поддерживающей REST, в которой отдельные базовые сервисы могут быть представлены:

- встраиваемыми компонентами в структуре многокомпонентных обработчиков текста,
- самостоятельными независимыми обработчиками текста, запрашиваемыми с произвольных клиентских систем (мобильные устройства, ноутбуки и нетбуки, стационарные персональные компьютеры).

Разработанный прототип платформы может быть применен в наукометрии и библиометрии при создания и развертывания Интернет-сервисов обработки и моделирования слабоструктурированных текстов статей, монографий, тезисов и целом литературных источников различной направленности для выявления скрытых свойств текстовых документов и междокументальных связей, что обеспечит еще более тесную интеграцию исследовательских групп в едином информационном пространстве, а также будет способствовать развитию связей образования, науки и бизнеса.

ЛИТЕРАТУРА:

1. Apache pdfBox documentation [Электронный ресурс] /Apache Group// Режим доступа: URL http://pdfbox.apache.org/index.html (Дата обращения: 30.01.2017 г.).
2. Buckley New retrieval approaches using SMART/ Buckley, Chris, Amit Singhal, Mandar Mitra, and Gerard Salton//TREC 4. In D.K. Harman (ed.) .
3. Peter H., Sach H., Bechstein C. Smartindexer — Amalagamating Ontologies and Lexical Resources for document indexing // In Proceedings of OntoLex- 2006. 2006.
4. Proceedings of the Eighth Global WordNet Conference [Электронный ресурс] / Editors: Verginica Barbu Mititelu, Corina Forăscu, Christiane Fellbaum, Piek Vossen.
5. Жевнерчук Д.В., Суркова А.С., Цыбульская Е.В., Чернобаев И.Д. RESTful-сервис обработки текстов // Материалы XXIII Международной научно-технической конференции «Информационные системы и технологии» (ИСТ-2017). 2017. — С. 348–352.
6. Жевнерчук Д.В. Моделирование процессов самоорганизации распределенных пространственно-временных ресурсов в открытых вычислительных системах [Текст] / Д.В. Жевнерчук //Вестник Нижегородского университета им. Лобачевского: № 2(1) — Нижний Новгород: ННГУ им. Лобачевского, 2014. — С. 218–222.
7. Жевнерчук Д.В. Открытая сервис-ориентированная платформа для интеграции информационных систем [Электронный ресурс] / Д.В. Жевнерчук // 18-я Международная научно-техническая конференция «Информационные системы и технологии» ИСТ-2012: Сб. трудов международной научно-технической конференции — Нижний Новгород: НГТУ им. Алексеева, 2012. — 1 электрон. опт. диск (CD-ROM). — С. 117–119.
8. Налимов В.В. Наукометрия. Изучение развития науки как информационного процесса [Текст] / В.В. Налимов, З.М. Мульченко // М: Издательство «Наука».
9. Ссылки на онлайн инструменты анализа и обработки текстов (для английского языка) [Электронный ресурс].

10. Ссылки на онлайн словари [Электронный ресурс]
11. Цветкова В.А. Система научной и технической информации для современной России: строим заново или учитываем имеющийся опыт [Текст] /В.А. Цветкова, Р.С. Гиляревский, И.И. Родионов //Информационные ресурсы России.
12. Цветкова В.А. Научная библиотека в едином информационном пространстве институтов социальной памяти [Электронный ресурс] / В.А. Цветкова, Е.В. Кочукова // Культура: теория и практика.

REFERENCE:

1. Apache pdfBox documentation [EHlektronnyj resurs] /Apache Group// Rezhim dostupa: URL: http://pdfbox.apache.org/index.html (Data obrashcheniya: 30.01.2017).
2. Buckley New retrieval approaches using SMART/ Buckley, Chris, Amit Singhal, Mandar Mitra, and Gerard Salton//TREC 4. In D.K. Harman (ed.).
3. Peter H., Sach H., Bechstein C. Smartindexer — Amalagamating Ontologies and Lexical Resources for document indexing // In Proceedings of OntoLex-2006. 2006.
4. Proceedings of the Eighth Global WordNet Conference [Электронный ресурс] / Editors: Verginica Barbu Mititelu, Corina Forăscu, Christiane Fellbaum, Piek Vossen.
5. Zhevnerchuk D.V., Surkova A.S., Cybul'skaya E.V., CHernobaev I.D. RESTful-servis obrabotki tekstov // Materialy XXIII Mezhdunarodnoj nauchno-tekhnicheskoj konferencii «Informacionnye sistemy i tekhnologii» (IST-2017). 2017. — S. 348–352.
6. Zhevnerchuk D.V., Modelirovanie processov samoorganizacii raspredelennyh prostranstvenno-vremennyh resursov v otkrytyh vychislitel'nyh sistemah [Tekst] / D.V. ZHevnerchuk //Vestnik Nizhegorodskogo universiteta im. Lobachevskogo: № 2(1) — Nizhnij Novgorod: NNGU im. Lobachevskogo, 2014. — S. 218–222.
7. Zhevnerchuk D.V., Otkrytaya servis-orientirovannaya platforma dlya integracii informacionnyh sistem [EHlektronnyj resurs] / D.V. ZHevnerchuk // 18-ya Mezhdunarodnaya nauchno-tekhnicheskaya konferenciya «Informacionnye sistemy i tekhnologii» IST-2012: Sb. trudov mezhdunarodnoj nauchno-tekhnicheskoj konferencii — Nizhnij Novgorod: NGTU im. Alekseeva, 2012. — 1 ehlektron. opt. disk (CD-ROM). — S. 117–119.
8. Nalimov V.V., Naukometriya. Izuchenie razvitiya nauki kak informacionnogo processa [Tekst] / V.V. Nalimov, Z.M. Mul'chenko // M: Izdatel'stvo «Nauka». — 1969. — C. 192.
9. Ssylki na onlajn instrumenty analiza i obrabotki tekstov (dlya anglijskogo yazyka) [EHlektronnyj resurs].
10. Ssylki na onlajn slovari [EHlektronnyj resurs]
11. Cvetkova V.A., Sistema nauchnoj i tekhnicheskoj informacii dlya sovremennoj Rossii: stroim zanovo ili uchityvaem imeyushchijsya opyt [Tekst] /V.A. Cvetkova, R.S. Gilyarevskij, I.I. Rodionov //Informacionnye resursy Rossii. — 2016. — №2. — C. 2–8.
12. Cvetkova V.A., Nauchnaya biblioteka v edinom informacionnom prostranstve institutov social'noj pamyati [EHlektronnyj resurs] / V.A. Cvetkova, E.V. Kochukova // Kul'tura: teoriya i praktika. — 2017. — №1(16). — Rezhim dostupa: URL: http://theoryofculture.ru/issues/70/920/ (Data obrashcheniya: 27.08.2017).