Information Services & Use 38 (2018) 61–65
DOI 10.3233/ISU-180005
IOS Press

Transforming 50 years of data: A machine learning approach to create new revenue streams for traditional publishers

Usha B. Biradar*, Lokanath Khamari and Shrey Bhate
Molecular Connections Pvt. Ltd., 5, Brigade Seshamahal, Vani Vilas Road, Basavanagudi, Bengaluru 560004, Karnataka, India

Abstract. Content-to-data is a global trend in the Information Community and publishers are at the heart of this transition. Secondary Publishers (database producers) face strong headwinds in the building of revenue growth, but at the same time sit on massive amounts of data that has been curated and indexed over decades. Big Data technologies such as Machine Learning, Artificial Intelligence, and linked data are finally coming of age to help create new services for both primary and secondary publishers, unlocking new meaning and relevance to both open data, and proprietary content assets. This paper provides an insight into the proprietary technologies developed at Molecular Connections and how these can be leveraged to generate a new lease of life for services in a maturing secondary publishing market.

Keywords: Artificial intelligence, machine learning, machine aided abstracting and indexing, big data, linked data stores, scalability and adaptability

1. Status quo of abstracting and indexing in scholarly publishing

Contrary to the popular belief, the advent of Internet of Things, the Big Data revolution, and digitization have necessitated a fresh outlook to the extant need and relevancy of Abstracting and Indexing (A&I) services in scholarly publishing. For this new generation of information users/consumers the ability to balance staying current and updated on one hand while sifting through humongous amounts of data, ingesting only the relevant data in a minimalistic fashion on the other, is largely dependent on the ease of access to platforms that allow him/her to do that in real time with some prompts here and there about the qualitative and quantitative state of the content. Google and other state-of-the-art platforms allow for this sort of direct user-to-content interactions. There exists something for all levels of users on such platforms, ranging from simple keyword search to advanced recommendations. These platforms act as primary facets over content search that will eventually lead to specific publisher sites/platforms [1]. More often than not, however, an intermediate-to-basic user is either over flooded with a lot of low-relevancy results or is restricted by very few generic results on such ease-of-access platforms. Access to deep-indexed content and custom reports/summaries for sales and customer insights driving decisions are the need of

*Corresponding author. E-mail: usha.bb@molecularconnections.com.

hour for today's information seeker. These needs translate to having services in place that enable high-throughput content enrichment, knowledge discovery systems, and an entire knowledge ecosystem that caters to different discoverability needs across an entire body of legacy content together with the constant inflow of new content.

With stricter budgetary constraints for libraries and lack of cost-effective access to technological advancements in the case of publisher platforms, bridging the gap between content churn-out rate and meeting demands of knowledge discovery is a growing concern [2].

2. Exploiting the untapped potentials of machine learning and artificial intelligence

Although a lot of progress has been made in terms of the applications of Machine Learning (ML) and Artificial Intelligence (AI) for knowledge discovery tasks, it remains a very domain-dependent, time-consuming, and costly attempt at a solution [3]. If the core modules are not sufficiently flexible with changing variables, cost and time parameters can skyrocket and scaling becomes increasingly pricey. To ensure return on investments, at the very least, the Machine Learning and Artificial Intelligence modules must be scalable with a flexibility to remove/add/change a few parameters/inputs at a given point.

At Molecular Connections, we have built proprietary cartridges of ML and AI modules that are capable of catering to complex tasks such as domain indexing, topic or subject area classifications, named-entity extractions and recognitions, mining for associations, fast-track ontology creations, entity disambiguation, question and answering, and text and image parsers. All of these modules are plug-and-play and can be integrated seamlessly into any existing workflow solution or they can be assembled end-to-end to create a custom solution for a specific need. In addition to being flexible, these individual modules are re-trainable with new inputs and learning systems ranging from active to reinforcement learning. The detailed features and capabilities of these proprietary modules are listed in the following sections.

3. MC proprietary modules

MC PARSER^{TM}: A proprietary solution to process unstructured content and XMLs to a standard format. This module acts as a central ingestion module enabling information extraction from a broad spectrum of source data, normalizing the records for downstream processing, and packing records in reflections of acceptable standards for delivery in any content enrichment/indexing tasks. Features and components include:

- XML parsers: XML parsers consist of a library of parsers based on xml formats of all major publishers
- Non-XML parsers: Appropriate Non-XML parser for extracting text from PDF, doc and other formats as required
- OCR processing: Hard copies and archived source data are accommodated and processed in the workflow as digital OCR elements
- Transformation and standardization to a common format (For instance, NISO JATS)
- Additional conversions to standard formats; e.g., XMLs to PDFs
- Plug-and-play reconditioning
- Versioning of documents
- Validation and updating of schema

MC MINER™: A proprietary, high accuracy text mining solution. Features include:

- Plug and play rules/Scope definitions: Use an ontology/thesaurus for the classification. System is customizable for any number of rules with or without precedence (Lexical and Contextual rules)
- Complete Machine Learning Modules for classifications and topic modeling
- Ensemble named entity taggers: Dictionary (Multiple thesauri/Controlled Vocabularies) + Machine trained models + Rules
- Accompanying APIs for each subtask listed above
- Re-indexing with threshold criteria along with identifying archival content being affected by current changes in the model and/or heuristics
- Augmentation/Enhancement of existing indexing with versioning information
- Integrated feedback ingestion and learning system ensuring validations, Quality Monitoring, and Optimization

MC LEXICON™: A proprietary ontology ingestion, maintenance and workflow management system. Features Include:

- Flexible schema creation and customizable properties and specifications of concepts and relationships
- Plug-and-play modules for normalization of concepts with external resources.
- Role-based user management.
- Quality control pipelines with a range of validations at schema and data level and trace back mechanisms with user logs and reports
- Visual summaries.
- Bulk uploads via simple text files, excels, etc.
- SKOS-JSON-based format standards that alleviates the need for complete transformation across different standards
- APIs for Concepts, concept details, and properties and concept hierarchies

MC IDENTIFY™: A proprietary author and institution disambiguation and standardization system, enabling metadata standardization against an accepted standard. This module includes the complete normalization and benchmarking essentials prior to standards and allows for multi-standards usage. Features Include:

- A parser to ingest various forms of inputs
- A specifically-tuned MC Miner™ module to identify different categories of named entities (authors, organizations, people, geo-locations, etc.)
- Ontologies/CVs of named entities (authors, topics, institutions, places, etc., with standardizations)
- An AI-based clustering algorithm for automated clustering of similar named entities
- A Machine Learning rule-based engine to influence the said clustering via plug and play user inputs
- Allowing for manual intervention to disambiguate semi-automated data points
- APIs talking to the linked data store

MC ANALYSE™: An interactive tool for performing data analytics and visualization primarily focused on business intelligence. Features Include:

- Schema less, can be configured for various data formats
- Agile environment and configurable
- Supports multiple columns

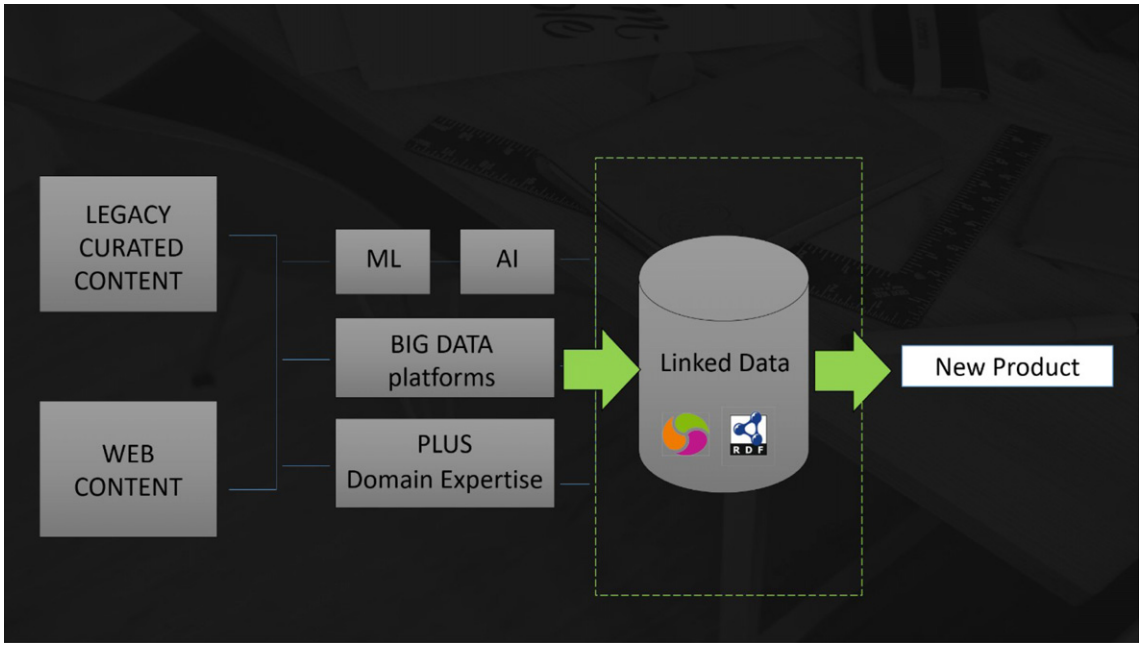

Fig. 1. New revenue model created by leveraging machine learning, artificial intelligence, and semantic data stores.

- Custom report formats can be produced
- Search and analytics
- Custom report generation
- Supports external API integration

4. Leveraging semantic/linked data stores to create and sustain new revenue models

As a result of using one or more of the modules described above, high-throughput content can be mirrored into a new, all-encompassing linked-entity model. A linked data store with its cellular components ranging from entities, attributes that define those entities, and the interactions amongst those entities is then weaved into a complex, but unique, knowledge ecosystem (KE). This KE then becomes the hub of the information flow on which cognitive tasks are applied to infer and analyze the knowledge flow (Depicted in Fig. 1). Key components that affect how this knowledge flow is perceived are heuristics, standards of the cognitive systems used (Reusability and reproducibility) and robustness/flexibility (both defined and redefined by spatio-temporal stimulants).

5. Conclusion

Publishing community's need of the hour is to be able to apply Machine Learning and Artificial Intelligence modules in order to solve complex problems and enable discoveries, ultimately creating knowledge ecosystems out of the voluminous data that they have already accumulated along with the additional data continuously being added. With the solutions/modules described in this paper, these

modern technological advances can be seamlessly employed, scaled-up, and sustained to create new revenue models for both primary and secondary publishers alike.

About Molecular Connections

Molecular Connections Pvt Ltd., was founded in 2001 by Jignesh Bhate upon whose presentation at the 2018 NFAIS Conference this paper is partially based. Bhate, its CEO, led the company from a start-up to a globally-respected organization that is the largest STM Indexing and Abstracting Company from India, with expertise that covers machine learning, text mining, literature curation, ontology development, content analytics, and visualization (see: http://www.molecularconnections.com/?page_id=18182). The corresponding author for this paper is Usha B. Biradar, a Decision Science Specialist at Molecular Connections. Email: usha.bb@molecularconnections.com.

References

[1] R. Van Noorden, November 07, 2014. "Google Scholar pioneer on search engine's future. As the search engine approaches its 10th birthday, *Nature* speaks to the co-creator of Google Scholar," *Nature News*, Nov. 7, 2017, Available from: https://www.nature.com/news/google-scholar-pioneer-on-search-engine-s-future-1.16269 (last accessed June 16, 2015).

[2] V. Larivière, S. Haustein and P. Mongeon, The oligopoly of academic publishers in the digital era, *PloS ONE*, June 10, 2015, http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0127502 (last accessed June 16, 2018).

[3] M. Ware and M. Mabe, *The STM Report: An Overview of Scientific and Scholarly Journal Publishing*, Fourth Edition, March 2015, pp. 146–151, https://www.stm-assoc.org/2015_02_20_STM_Report_2015.pdf, last accessed June 17, 2018.