

Aunar recursos de bibliotecas y museos: ¿cómo puede ayudar la inteligencia artificial, basándonos en el Proyecto de la Colección de Libros de Ivan Tsvetaev?

Spanish translation of the original paper: “Bringing Library and Museum Resources Together: How Can Artificial Intelligence Help (based on the Ivan Tsvetaev’s Book Collection Project)?”.

Translated by: Iván Pérez Marinas, Biblioteca Nacional de España, Madrid, España

El texto de este documento se ha traducido al español y puede presentar diferencias respecto al texto original. La presente traducción se aporta exclusivamente para fines de referencia.

Ekaterina Igoshina

Biblioteca de Investigación, Museo Estatal de Bellas Artes Pushkin, Moscú, Rusia
ekaterina.igoshina@arts-museum.ru.

Juliya Dubrovskaya

Biblioteca de Investigación, Museo Estatal de Bellas Artes Pushkin, Moscú, Rusia
julia.dubrovskaya@arts-museum.ru



This is a Spanish translation of “*Bringing Library and Museum Resources Together: How Can Artificial Intelligence Help (based on the Ivan Tsvetaev’s Book Collection Project)?*” copyright © 2019 by **Iván Pérez Marinas**. This work is made available under the terms of the Creative Commons Attribution 4.0 Unported License:
<http://creativecommons.org/licenses/by/4.0>

Resumen:

Hace unos años en la Biblioteca de Investigación del Museo Estatal de Bellas Artes Pushkin comenzamos un programa de digitalización a largo plazo de las partes más valiosas y/o frágiles de nuestra colección de libros. Para empezar decidimos dividir este proceso en un número de programas más pequeños, cada uno dedicado a una colección diferente. Una de nuestras primeras y evidentes elecciones para digitalizar, y así hacerla visible y totalmente accesible, fue la colección del fundador y primer director de nuestro museo, Ivan Tsvetaev (1847-1913), que está formada por más de 850 volúmenes y recalca vivamente sus intereses científicos, lazos académicos y aficiones. Mientras digitalizábamos los libros de Tsvetaev descubrimos un número bastante grande de dedicatorias, casi 250; una fuente complementaria de información muy útil sobre el ámbito científico y civil de nuestro fundador. Todas las inscripciones se escanearon, se descodificaron y, cuando hacía falta, se tradujeron al ruso. Cada una de las imágenes de las inscripciones fue también enriquecida con breves biografías de los autores de las inscripciones (donantes).

En primer lugar, que además era el paso más evidente, reunimos nuestros diversos recursos digitales en un espacio semántico común, una colección digital, usando herramientas técnicas que estaban a nuestra disposición. Al compilar las descripciones de todas las unidades relativas a la colección en nuestro catálogo y al enlazarlas con nuestro sistema de datos de autoridad unificado, como los índices de materia y persona, logramos crear un fragmento semántico muy cómodo de usar que describe mejor la personalidad de Ivan Tsvetaev y la historia de la creación de nuestro museo a través del prisma de sus intereses personales y contactos profesionales.

Entonces empezamos a contemplar la vía de una posible conexión de esta colección a otros recursos museísticos complementarios, que están guardados en otros departamentos y descritos en otros sistemas de tecnología de la información (TI). Uno de los posibles enfoques para aunar todas estas fuentes no coordinadas es aplicar tecnología de inteligencia artificial. Hallamos esta manera intrigante y bastante prometedora.

Nuestro artículo está dirigido a explicar la posible vía de relacionar datos para diseñar un diálogo entre el asistente de voz y el usuario mientras busca en una biblioteca, un museo y unos posibles recursos externos, basándonos en el ejemplo de una determinada colección digital.

Palabras clave: biblioteca de museo, colecciones digitales, relación de datos, tecnología de inteligencia artificial, asistente de voz.

Introducción

Hace unos años comenzamos un programa de digitalización de larga duración dirigido a preservar del deterioro las partes más valiosas y/o frágiles de la colección de libros de la Biblioteca de Investigación del Museo Estatal de Bellas Artes Pushkin. Para dar una idea general de las posesiones de nuestra Biblioteca para todos aquellos que no lo conocen nos gustaría señalar las siguientes cuestiones:

- nuestra Biblioteca abrió sus puertas al público el 31 de mayo de 1912 (el mismo día que fue inaugurado todo el Museo);
- al mismo tiempo de la apertura y durante las dos primeras dos décadas de su existencia la Biblioteca solía servir como un importante instrumento auxiliar para los investigadores del museo y los conservadores de la colección y como la biblioteca de arte de referencia pública para usuarios externos (la mayoría de ellos eran visitantes habituales del museo);
- el núcleo de su colección se origina a partir del Gabinete de Bellas Artes del Departamento de Filología de la Universidad de Moscú del que el fundador del Museo, Ivan Tsvetaev, era responsable antes de convertirse en 1912 en el primer director del museo recién fundado;
- actualmente nuestra colección comprende unas 220.000 unidades de almacenamiento que incluyen libros, folletos y publicaciones periódicas que cubren toda la historia del arte occidental desde la antigüedad hasta nuestros días;
- casi 40.000 unidades de almacenamiento del conjunto de la colección son los llamados materiales raros y valiosos, tanto libros como publicaciones periódicas, que necesitan de un cuidado especial y son evidentemente los primeros en ser sujetos del programa de digitalización.

Para empezar, siguiendo las prácticas de digitalización más exitosas del mundo, decidimos dividir el proceso en un número de programas más pequeños que cubrieran colecciones diferentes. Una de nuestras primeras y más evidentes elecciones para digitalizar, para así hacerlo visible y contribuir a un acceso público mayor, fue la colección del susodicho fundador del museo y primer director, Ivan Tsvetaev (1847-1913). Fue donada a la Biblioteca del Museo de Bellas Artes Emperador Alejandro III, futura Biblioteca de Investigación del Museo Estatal de Bellas Artes Pushkin, tras la muerte de Tsvetaev en 1913.

La colección de libros de Tsvetaev consiste en más de 860 volúmenes y recalca vivamente sus intereses científicos, lazos académicos y aficiones. Debido al interés del propietario tanto por las lenguas y arqueología latinas antiguas como por la historia del arte occidental desde la antigüedad hasta finales del siglo XVI, la mayoría de sus libros está dedicada a estos dos temas. Como era profesor universitario nunca tuvo suficientes recursos financieros para comprar libros realmente raros o con buenas ilustraciones ni para encuadernar sus libros de una manera espléndida. A pesar de que la mayor parte consiste en publicaciones académicas con encuadernaciones relativamente modestas, para nosotros presenta una rica fuente de información histórica sobre el proceso de planificación, construcción y finalmente apertura de nuestro Museo, además de sobre su persona, cuyo papel en la fundación de nuestro Museo ha sido reconocido como crucial.

Mientras digitalizábamos los libros de Tsvetaev descubrimos un número bastante grande de dedicatorias, casi 250; una fuente complementaria de información muy útil sobre el ámbito científico y civil de nuestro fundador. Una parte importante de las inscripciones representaba dedicatorias simples y bastante formales, mientras que otras contenían textos bastante largos que hacían referencia al tipo especial de relaciones que le unieron con sus autores. Entre los que habían dejado sus caligrafías en los libros de Tsvetaev había numerosos colegas suyos (profesores de filología, arqueología e historia del arte procedentes de Rusia y todo el mundo), estudiantes de su antigua universidad y algunos amigos cercanos de la familia. Todas las dedicatorias que se descubrieron en los libros de Tsvetaev fueron escaneadas con cuidado, se descodificaron y, cuando hacía falta, se tradujeron al ruso. Cada una de las imágenes de las inscripciones fue también provista de breves biografías de sus autores, o donantes.

En primer lugar, que además era el paso más evidente, integramos nuestros diversos recursos digitales en un espacio semántico común dentro del entorno software de la Biblioteca, la colección digital; lo hicimos con la ayuda de herramientas técnicas que estaban a nuestra disposición. Al compilar descripciones de todas las unidades relativas a la colección en nuestro catálogo y al enlazarlas con nuestro sistema de datos de autoridad unificado, como los índices de materia y persona, logramos crear un fragmento semántico bastante fácil de usar que describe mejor la personalidad de Ivan Tsvetaev y la historia de la fundación de nuestro Museo a través del prisma de sus intereses personales y contactos profesionales.

Entonces empezamos a trabajar en el problema de proveer una información más completa sobre enlaces históricos que percibimos al descodificar las inscripciones de los donantes. Es muy valioso al estudiar fuentes, ya que contribuye a establecer vínculos y referencias cruzadas entre esta colección y otras fuentes relevantes del museo que están guardadas en otros departamentos y están basadas en un sistema de tecnología de la información (TI) diferente. Pero el problema es que incluso dentro de una misma institución puede no haber una solución técnica para desarrollar la cohesión semántica de los objetos del museo pertenecientes a departamentos diferentes, por no mencionar los vínculos a fuentes externas.

Por ejemplo, hay dos pequeños folletos entre los libros de Tsvetaev donados por su antiguo estudiante Nikolay Romanov muy a principios del siglo XX. Ambos folletos contienen unas inscripciones casi idénticas y muy cortas: “A mi querido y profundamente respetado profesor Ivan Vladimirovich Tsvetaev de parte de su devoto autor”. Nikolay Romanov, uno de los más brillantes estudiantes de Tsvetaev, graduado por la Universidad de Moscú en 1890, en 1898 entró a formar parte del Comité de la Fundación del Museo de Bellas Artes, convirtiéndose así en colega y el asistente más leal de Tsvetaev. En 1923, diez años después de la muerte de Tsvetaev, le sucedió ocupando el cargo de director del Museo y permaneciendo en él hasta 1928. Navegando por el catálogo de nuestro Museo (una base de datos especial para todo tipo de objetos del museo compilados por nuestros conservadores de colecciones de diferentes departamentos), se pueden encontrar fácilmente más de 20 objetos del museo relacionados con Nikolay Romanov: algunas de sus fotos, 10 retratos gráficos en diversas técnicas, el escaneo de su publicación periódica sobre la historia del Museo, así como su exlibris. Al mostrar este ejemplo nos gustaría remarcar la idea de que varios recursos asociados que están identificados por separado en los sistemas TI de la Biblioteca y del Museo tienden realmente a complementarse recíprocamente y crear un espacio temático dedicado a una persona determinada, pero en la práctica no cumplen este objetivo cuando se buscan por separado.

Sopesando varias alternativas avanzadas que ofrece el mercado de TI actual en el área de la búsqueda integrada, nos hemos interesado en la solución de vanguardia que ha logrado un gran avance en los últimos dos años: las tecnologías de inteligencia artificial.

Cuando usamos este término, “inteligencia artificial”, nos referimos al tipo de aplicaciones que son capaces de simular ciertos procesos de inteligencia humana usando una interpolación matemática compleja. Por ejemplo, son capaces de reconocer imágenes visuales, palabras habladas, tomar decisiones con la ayuda de algoritmos y desarrollar tendencias basándose en análisis estadístico. Algunos programas de soluciones de inteligencia artificial (IA) se están volviendo fácilmente accesibles en este momento. Los motores de búsqueda más grandes abren sus bibliotecas de programas, aplicaciones de reconocimiento de voz y algoritmos de aprendizaje automático.

Las tecnologías de reconocimiento de voz que se llevaron a cabo por medio de los llamados asistentes de voz o asistentes inteligentes fueron bien recibidas y ahora están obteniendo rápidamente reconocimiento público. Según las previsiones de las agencias de análisis, alrededor del 2020 entre el 30% y el 50% de las consultas de búsqueda en internet se realizarán mediante petición verbal. Y hay una explicación razonable a esto. La habilidad del asistente de voz es una nueva clase de trabajo con bases de datos. El usuario recibe la información solicitada mientras lleva una conversación natural. Permite al usuario mantener su ritmo diario y administra la base de datos de la manera que corresponde a los requisitos ergonómicos modernos de las tecnologías, al mismo tiempo que lleva la relación calidad-cantidad de los resultados de búsqueda al grado más adecuado para la percepción humana.

La consulta de voz no requiere por parte del usuario de un conocimiento sobre sintaxis de consultas de búsqueda especial o sobre estándares de metadatos que describen los recursos. La pregunta puede formularse de cualquier forma que el usuario estime conveniente en un lenguaje humano simple. Su propia característica de asistente de voz ya las convierte en un asunto de seria consideración para introducir esta tecnología en la búsqueda bibliográfica.

Las tecnologías de reconocimiento de voz dividen la frase en palabras, definiendo el *script* de búsqueda especialmente desarrollado los dominantes semánticos y, finalmente, las variantes de las consultas de los usuarios enviadas se transforman en las llamadas frases clave que

guiarán nuevas búsquedas. Y merece la pena mencionar que el número de tales frases reconocidas está creciendo constantemente; la inteligencia artificial “aprende” de cada consulta.

Aparte de eso, la característica importante de la búsqueda con asistente inteligente es la oportunidad de interactuar con el usuario, que no se queda solo enfrentándose con los resultados de búsqueda, sino que se le ofrece una continuación del diálogo por medio de la conversación de búsqueda. Así hay una oportunidad que permite al usuario conseguir resultados más satisfactorios, y más que eso, hay una oportunidad de que reciba respuesta sobre la calidad de la consulta realizada.

Sin embargo, las oportunidades susodichas no abarcan todas las posibles ventajas de usar tecnologías IA mientras se procesan bases de datos bibliográficas y recursos de información electrónica. La tecnología de reconocimiento de voz contiene potentes algoritmos del llamado “aprendizaje automático”. Las bibliotecas de software que llevan a cabo estos algoritmos están siendo constantemente actualizadas y desarrolladas por los desarrolladores más importantes del mundo en el área de la tecnología cognitiva y están abiertas a acceso público.

Los algoritmos de aprendizaje automático son de hecho la base de la tecnología de la inteligencia artificial. Proporcionan extracción de reglas, funciones, dependencias en los elementos de datos basadas en modelos matemáticos y principios de la red neuronal. Y si en el pasado el desarrollo del algoritmo de búsqueda de la base de datos estaba completamente programado por el ser humano, y los datos exigían preparación preliminar, es decir, formalización a este algoritmo, hoy en día los algoritmos de aprendizaje automático pueden analizar tanto el contenido de la base de datos como el autoestudio sobre las muestras de resultados exitosos de consultas de búsqueda al encontrar la interrelación entre la estructura de la base de datos que se refiere a la consulta del usuario y la consulta misma. Así, cuanto mayor sea el número de consultas realizadas que reciba el algoritmo de aprendizaje automático, más exacto será el modelo de estructura de búsqueda sobre la posible fuente que podrá llevar a cabo.

Por lo tanto, en teoría los algoritmos de aprendizaje automático serán capaces de hacer frente al problema de precisión en seguir los estándares mientras se preparan los metadatos.

Lamentablemente debemos tener en cuenta que, a pesar de los estándares desarrollados para definir los elementos de la base de datos, puede haber varias formas de interpretarlos, e incluso pueden variar enormemente los volúmenes de una misma colección multi-volumen definida en la base bibliográfica por varias personas en varios años, de modo que los algoritmos de búsqueda clásicos no podrán encontrar ambas descripciones en una misma consulta de búsqueda.

Integrar bases de datos para seguir con precisión los estándares a menudo es imposible de lograr debido principalmente a la cantidad de datos sujetos a reescritura.

El resultado de la implementación de tecnologías de IA en la búsqueda a través de las bases de datos del museo se puede representar en la siguiente tabla:

Lo que hacemos ahora: separar la búsqueda a través de varias bases de datos.

		Búsqueda en	Sistema de información automatizada (SIA) de la Biblioteca
		Búsqueda en	SIA del Museo
		Búsqueda en	SIA del Archivo
		Búsqueda en	SIA externo

Lo que puede hacerse usando aplicaciones de IA*: búsqueda integrada

Voz	Intelectual	Búsqueda en	SIA de la Biblioteca Museo Archivo Conjunto de datos externos
- <u>Reconocimiento de voz por IA*</u> , transmisión de palabras articuladas en la búsqueda - El <i>script</i> del diálogo con el usuario (qué buscar, dónde buscar, cómo enviar los resultados)	- El <i>script</i> de búsqueda en la bibliografía usando <u>bibliotecas de aprendizaje automático por IA*</u> para el análisis de texto sin tener en cuenta las reglas de lectura del formato, basándose únicamente en las consultas cumplidas y su estimación - El sistema de registro de acciones del usuario como modelo de aprendizaje (acciones del usuario, consultas relacionadas) - La base de datos de las mejores respuestas a consultas estándar		Integrando varios recursos de búsqueda en un único almacén en formato JSON (en la tabla), incluyendo datos abiertos enlazados externos publicados y otros tipos de conjuntos de datos.
En un lenguaje natural simple	Búsqueda en el texto sin tener en cuenta el estándar de entrada de datos y su cumplimiento		Búsqueda cruzada

*Los lugares donde se aplican tecnologías IA están señalados con subrayado

Hasta el momento no tenemos noticia de ejemplos de sistemas de IA tan ampliamente difundidos en las bases de datos bibliográficas, pero creemos que implementar tecnologías de IA en las bases de datos de los museos dará como resultado un sistema de autoaprendizaje unificado que será suficientemente flexible y hablará lenguaje natural, omitiendo la falta de homogeneidad en datos y entregando resultados bastante adecuados sin necesidad de un procesamiento preliminar suplementario de datos.

Merece la pena mencionar que, mientras le trabajamos en resolver el problema de la inexactitud siguiendo los estándares de descripción del contenido de las colecciones, no afirmamos en absoluto que la estandarización ya no sea un problema real. Las tecnologías de

IA, desde nuestro punto de vista, son sólo un medio universal y flexible para producir datos que no contraviene los estándares de metadatos pero que hace que el impacto de las inconsistencias sea menos crucial para los resultados de búsqueda. Siguiendo los estándares de metadatos podemos conectar los datos entre sí, pero presentarlos al usuario final requiere de una propuesta fácil de usar capaz de realizar búsquedas en todas partes donde sea posible y entregar el resultado en términos claros.

La inteligencia artificial capaz de reconocer varios tipos de conjuntos de datos y aplicar una sintaxis adecuada para extraerlos puede reemplazar el software especial que funciona con un único algoritmo y formato de datos.

Y volvamos a nuestro Proyecto de Colección de Libros de Ivan Tsvetaev, que en nuestro caso se convirtió en un modélico fragmento de los datos para desarrollar el modelo de datos para el asistente de voz.

En marzo de 2018 Yandex, el famoso sistema de motor de búsqueda ruso, lanzó para pruebas beta la plataforma Yandex Dialogues. Esta plataforma da libre acceso al software de reconocimiento de voz implementado por medio del asistente inteligente Alice. Así todos los desarrolladores de software pueden utilizar este software en sus propios proyectos. Alice es el equivalente ruso del Asistente de Google, Amazon Alexa o Windows Cortana. Lo encontramos tanto desafiante como ambicioso.

El uso de este tipo de asistente de voz requería la escritura preliminar de un pequeño programa, la llamada "habilidad" que contenía el *script* de conversación con el usuario. En el marco de nuestro experimento modélico combinamos dos tipos de datos: los datos de la base bibliográfica y las descripciones de los objetos tipificados por una persona relativos a la colección que proceden de los registros contables del museo. Al desarrollar esta "habilidad" llegamos a comprender el hecho de que los asuntos más importantes de todo son predecir y modelar las consultas del usuario al respecto y desarrollar el *script* de diálogo entre Alice y el usuario, y viceversa, entre el usuario y Alice.