DESARROLLO DEL ESTADO DEL ARTE EN INVESTIGACIÓN: UNA HERRAMIENTA BASADA EN INTELIGENCIA ARTIFICIAL

Víctor Andrés Bucheli Guerrero

Doctor en Ingeniería, Profesor Facultad de Ingeniería, Escuela de Ingeniería de Sistemas y Computación, Grupo de Investigación GUIA, Universidad del Valle, Cali, Colombia. Correo electrónico: victor.bucheli@correounivalle.edu.co
Escuela de Ingeniería de Sistemas y Computación, Facultad de Ingeniería, Universidad del Valle, Calle 13 N.º 100-00,76001, Cali-Valle del Cauca, Colombia.

RESUMEN

El trabajo describe el proceso semi-asistido de desarrollo del estado del arte en investigación. Así, se presenta un prototipo de software basada en inteligencia artificial, un estudio de caso de un proyecto de doctorado y los resultados de una encuesta llevada a cabo a 40 estudiantes de maestría y doctorado, quienes reportan la utilidad de la herramienta. En el documento se presenta la arquitectura, la implementación del prototipo de software y un estudio comparativo con las herramientas existentes. Se discuten las potencialidades del proceso asistido por la herramienta y el impacto positivo que puede tener en la investigación, principalmente en el contexto colombiano. Así también se discuten las técnicas de inteligencia artificial implementadas, la escalabilidad de la herramienta y la facilidad de integrar nuevos análisis y visualizaciones.

Palabras clave: estado del arte, aprendizaje de máquina, visualización, reconocimiento de patrones, investigación.

Recibido: 19 de septiembre de 2019. Aceptado: 27 de Noviembre de 2019.
Received: September 19, 2019. Accepted: November 27, 2019.

Semi-automatic systematic reviews in scientific research: a software tool based on artificial intelligence

ABSTRACT

This paper describes a process of semi-automatic systematic reviews in scientific research. On the one hand, it is presented a software prototype based on artificial intelligence, semi-automatic systematic reviews tool. On the other hand, a study case of a phd thesis project shows the maps and information with aggregate value. In addition, the software prototype were evaluated by 40 students of master and Ph.D projects, they report the utility of the software tool by a survey. This work presents the modules and functionalities, software architecture and software deployment. Furthermore, a comparative study of a set of similar software tools has been conducted according to the following criteria: input files, analysis and mapping, and software license. Finally, it is discussed the results of the comparative study, the implementation of artificial intelligence techniques, the software scalability and the easiness of include new analysis and maps.

Keywords: *systematic reviews; machine learning; visualization; pattern recognition; scientific research*

Cómo citar este artículo: V. Bucheli. "Desarrollo del estado del arte en investigación: una herramienta basada en inteligencia artificial.", Revista Politécnica, vol. 15, no.30, pp.70-81, 2019. DOI: 10.33571/rpolitec.v15n30a7

1. INTRODUCCIÓN

El sistema apoya la construcción de una revisión sistemática de literatura. La revisión sistemática de literatura es el proceso por el cual se inicia un proyecto de investigación o la definición de un tema de investigación. La revisión sistemática tiene por objetivo identificar y describir el estado del arte y de la técnica. Por lo general existen metodologías para construir el estado del arte o la revisión sistemática de literatura, las cuales generalmente construyen un corpus de referencias bibliográficas que mapean el estado actual de desarrollo de la temática en cuestión [1]–[3].

Una revisión sistemática de literatura es un proceso de investigación de tipo documental [1], el cual permite definir un tema de investigación. El proceso parte de la definición de ecuaciones de búsqueda que se van refinando en la medida que se cuenta con más información sobre el tema por ejemplo palabras clave o autores relevantes. El proceso se basa en identificar referencias útiles, almacenar sistemáticamente las referencias y actualizar iterativamente la lista de referencias. Así las referencias encontradas se convierten en un insumo para hacer investigación en sí misma. Al resultado final del proceso se le conoce como corpus de referencias, éste es el insumo para definir cuáles son las preguntas abiertas de investigación en un tema de investigación particular. Según el diccionario de Oxford, un *systematic review* es el estado reciente de desarrollo de una tecnología, el cual incorpora ideas y características novedosas [4]. A su vez, el estado del arte puede ser entendido como una expresión que se refiere al desarrollo más reciente en un campo [5].

Este trabajo se centra en el proceso de construcción del corpus de referencias y la descripción de una herramienta que soporta el análisis de dicho corpus. En este sentido se entiende al corpus de referencias como una base de conocimiento que permite entender en profundidad el contexto y las características del tema de investigación o campo de investigación.El corpus es el punto de partida para desarrollar un análisis crítico que permita identificar preguntas de investigación abiertas o problemas de investigación.

El proceso es evaluado por medio de un estudio de caso en un proyecto de investigación de doctorado. De igual manera se realizó una encuesta a estudiantes de maestría y doctorado, quienes están iniciado con la formulación de sus proyectos de investigación y utilizaron la herramienta propuesta en este trabajo.

Actualmente existen diferentes herramientas que soportan el proceso de almacenamiento y organización de referencias bibliográficas, entre ellas Zotero [6], Mendeley [7], etc. Sin embargo, dichas herramientas no permiten hacer análisis sobre el desarrollo de la temática, pues el formato de almacenamiento es un texto marcado por etiquetas, formato de almacenamiento conocido como bibtex [8]. En este sentido hacer análisis del corpus de referencias no es posible en las herramientas existentes. Por ejemplo, analizar la dinámica de referencias por año, es una tarea que se debe realizar de forma manual. Existen otras aplicaciones desarrolladas para hacer *science mapping,* sin embargo estas son desarrolladas con propósitos diferentes, enfocadas al análisis bibliométrico o cienciométrico [9], [10].

Semi-Automatic Systematic Review (SASR) es una herramienta que soporta el análisis de corpus de referencias, el prototipo se desarrolló para el área de ingeniería de sistemas y computación, en este sentido, las clasificaciones utilizadas son propias del área, por ejemplo, la clasificación de la ACM (Association for Computing Machinery). La herramienta permite realizar análisis descriptivos, así como también, análisis más complejos como por ejemplo clustering de documentos.

La aplicación prototipo toma como fuente de datos un archivo bibtex, archivo que contiene el corpus de referencias, corpus que se ha construido mediante un proceso que permite identificar un tema de investigación. A partir de éste archivo la herramienta permite reconocer por medio de mapas la dinámica de una temática, identificar las palabras clave relevantes, los autores más referenciados en el campo, la red de coautoría, filtrar la información por palabras clave, año o autores, adicionalmente, la herramienta permite a su vez navegar por el corpus construido.

La herramienta prototipo SASR integra estadísticas descriptivas y técnicas de inteligencia artificial. Las técnicas de Inteligencia permiten entender la temática en cuestión por medio del reconocimiento automático de patrones y visualizaciones tales como treemap [11] o radial tidy tree [12]. Así SASR se convierte en un software novedoso para entender un corpus de referencias, permitiendo al investigador identificar, reconocer patrones, navegar y visualizar información acerca del tema de investigación. Así, la herramienta prototipo soporta el proceso de análisis de las referencias y la construcción de una revisión sistemática de literatura. La herramienta está disponible en [13] y el código fuente está disponible en [14].

El artículo está organizado de la siguiente manera: en la sección 2 se presentan trabajos relacionados, en la sección 3 se presenta la arquitectura del sistema SARS, seguido de la sección 4 en la que se presentan los resultados y finalmente la sección 5 la discusión.

2. TRABAJOS RELACIONADOS

El *tech mining* explora y explota la información científico tecnológica, buscando patrones para describir y entender un proceso de innovación tecnológica. Concretamente, el *tech mining* utiliza las herramientas de aprendizaje de máquina aplicadas a la documentación científica [15]. Las fases del proceso son las siguientes: selección de fuentes de información, recuperación de información relevante, limpieza de datos, análisis básicos, construcción de indicadores y mapas, interpretación y comunicación de los resultados. Las fuentes de información para estos análisis pueden ser principalmente papers académicos o patentes.

El *tech mining* se implementa en las empresas para apoyar la toma de decisiones: buscar nuevos mercados o la rentabilidad de una nueva tecnología; anticiparse a las trayectorias futuras del desarrollo de tecnologías e identificación de pasos evolutivos para el desarrollo de una tecnología; identificar socios o competidores potenciales; y planeación [15].

Otra área relacionada con el presente trabajo es *science mapping*, ésta permite mostrar a través de mapas bibliométricos las diversas disciplinas que constituyen un área de la ciencia, específicamente, la estructura conceptual, intelectual y social [9], [10]. El proceso que sigue generalmente para el mapeo científico es: recuperación de información, preprocesamiento, extracción de redes, normalización de la información, mapeo, análisis y visualización [16]. Las fuentes de información para estos análisis pueden ser papers académicos, patentes o datos de financiamiento de la investigación [17], [18]. De esta manera, ésta área se enfoca en monitorear un área de la ciencia y delimitarla para determinar su estructura de conocimiento y evolución [19].

En la revisión de literatura se encuentra software tanto para *tech mining* como para *science mapping*. Según [16] Bibexcel es un toolbox que permite llevar a cabo análisis bibliométricos tales como distribuciones de frecuencia de citaciones, autores, análisis de coocurrencias, entre otros. Bibexcel maneja diferentes tipos de datos, sin embargo, en [16] se reporta que los nuevos usuarios perciben que la herramienta no es fácil de utilizar. Adicional a ello la herramienta no permite hacer análisis más robustos o visualizaciones, para ello Bibexcel permite exportar los archivos a formatos de archivos de SPSS [20] o Pajek [21].

CiteSpace II es una herramienta de software para detectar, analizar y visualizar patrones y tendencias en literatura científica. El objetivo de la herramienta es analizar tendencias emergentes en la ciencia [22], [23]. CoPalRed es un software comercial que permite análisis de coaparición de palabras clave de los documentos científicos. Los archivos que recibe son csv o separados por comas. CoPalRed permite análisis de redes temáticas, red temática relativa frente al área de la ciencia y cambios en el tiempo de la red temática [24], [25].

VantagePoint es una herramienta de minería de texto, la herramienta es comercial, ésta es útil para descubrir conocimiento en patentes o documentos académicos, principalmente información textual. La herramienta permite analizar grandes volúmenes de texto y descubrir patrones enfocados en preguntas propias del *tech mining:* quién produce el conocimiento?, qué conocimiento produce?, cuándo y dónde se produce?.

VantagePoint es una herramienta robusta que permite la importación de distintos tipos de archivos, en la herramienta se navega por la información, así como también realizar análisis descriptivos, frecuencias, histogramas, y visualizaciones de relaciones tales como matriz de coocurrencia, matrices de autocorrelación y matrices de correlación cruzada. VantagePoint preprocesa y limpia los datos, encuentra automáticamente datos repetidos o candidatos a ser datos repetidos, también permite la utilización de tesauros. Finalmente, la herramienta permite la visualización de *cross-correlation map, auto-correlation map, and factor map* [15], [26], [27].

VOSViewer es un software útil para visualizar mapas bibliométricos de cualquier tipo de datos de coocurrencia [28], la herramienta de software no permite la construcción de las matrices sólo permite la visualización. Sin embargo en términos de visualización es una herramienta robusta, la herramienta permite label-view, density-view, cluster density-view y scatter view [10], [29].

Vigtech es una herramienta de soporte a la vigilancia tecnológica. A partir de un conjunto de documentos de SCOPUS la herramienta permite extraer metadatos, cálculo de estadísticas descriptivas, análisis de redes sociales, análisis de redes de palabras claves y visualización [30].

VigHub es una herramienta que soporta los procesos de vigilancia tecnológica y permite ser la fuente de información para la construcción de estados del arte relacionados con el desarrollo de software [31]. La herramienta toma los datos de la plataforma GitHub [32], a partir de la información que se extrae se construyen mapas tecnológicos que describen el estado y la evolución de proyectos de software y las perspectivas de tecnologías específicas. La herramienta permite obtener mapas tecnológicos tales como, lenguajes de programación o cronología de evolución de temas [31].

Este trabajo tiene como propósito mostrar una arquitectura del software SARS para la construcción semi-asistida del estado del arte en un proceso de investigación. En los trabajos revisados hay software similar, pero no se reporta en la literatura un software que asista el proceso de desarrollo del estado del arte en una investigación.

Los mapas obtenidos de la herramienta SASR permiten obtener información agregada, e información que no se conocía sobre el tema de investigación. De igual manera, las estadísticas que se muestran en SASR ayudan a clasificar los documentos, enfocar el tema e identificar principales autores. El mapa basado en *The ACM Computing Classification System (CCS)* y treemap permiten entender el área de investigación como un árbol de temas e identificar nuevos artículos significativos para la investigación.

3. MATERIALES Y MÉTODOS

El modelo de SASR parte de los trabajos de [30], [31]. El modelo de SASR caracteriza un corpus de referencias automáticamente, donde las referencias bibliográficas provienen de un archivo Bibtex. En SASR se implementan métodos y algoritmos de procesamiento y análisis de datos, los cuales extraen patrones y visualizaciones para entender el estado actual de un tema de investigación.

3.1 ARQUITECTURA SASR

La herramienta propuesta es un sistema modular caracterizado por su extensibilidad y escalabilidad vertical, con la posibilidad de integrar fácilmente nuevos desarrollos. El prototipo está basado en microservicios a través de contenedores, lo que se conoce como dockerización [33]. A continuación, se presenta la descripción de la arquitectura de componentes y lógica que permite la integración de diferentes técnicas computacionales implementadas en diferentes lenguajes de programación.

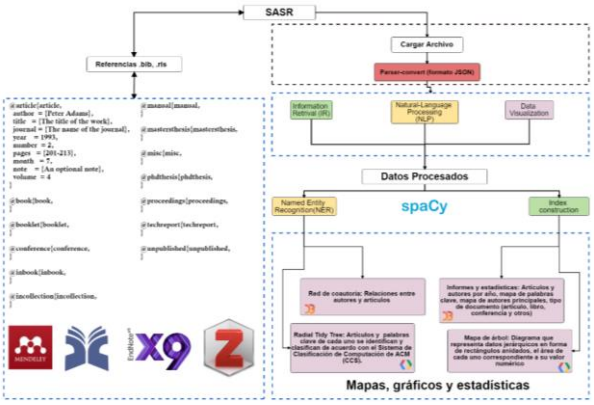

Fig.1. Arquitectura de componentes de SARS.

La Figura 1. presenta la arquitectura por componentes del prototipo, e indica el proceso secuencial en el que se ejecuta el sistema. Se propone una arquitectura híbrida donde se siguen principios de diseño de las arquitecturas orientadas a microservicios (*Model View Controller* MVC). La herramienta se implementó en el framework Flask [34].

Los microservicios son totalmente independientes y autocontenidos de modo que la ejecución se pueda realizar de forma aislada. Por su parte la Figura 2. muestra la arquitectura lógica donde se encuentran los módulos que se generaron en la construcción del prototipo.

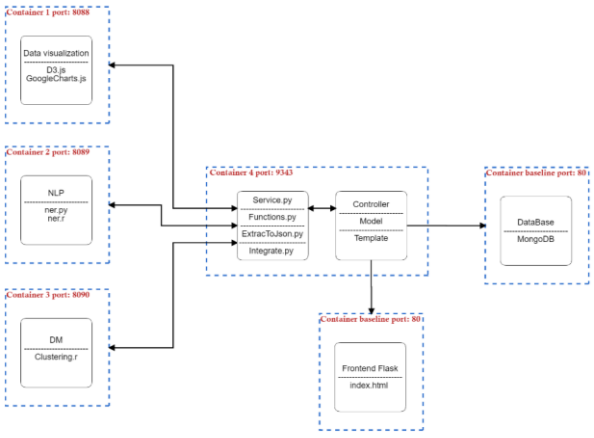

Fig.2. Arquitectura lógica de SARS.

La capa de vista propuesta para esta arquitectura utiliza el framework flask de Python[34], JavaScript [35], la librería Google Chart [36] y D3.js [37], para la interacción con el usuario y la visualización de datos. La capa vista está totalmente desacoplada del servidor, es decir, la lógica detrás de la renderización de la interfaz gráfica de usuario, no pasa por el

procesamiento del servidor, si no que se lleva en su totalidad en el navegador del usuario.

La capa modelo está compuesta por una base NoSQL, eso quiere decir que guarda estructuras de datos BSON, especificación similar a JSON. Acá se almacenan los archivos de los usuarios y sus credenciales, también se usa la base NoSQL para el almacenamiento de los metadatos extraídos de cada lista de referencias de forma independiente, esto es, la información del proyecto de un usuario se almacena en un JSON y un índice independiente, lo que permite el rápido acceso a la información semiestructurada. Esta información contiene datos relacionados a archivos, palabras claves, autores, año de la referencia, co-autores, entre otros datos que sirven de insumo a los módulos definidos en el controlador.

La capa del controlador consiste en un conjunto de módulos que siguen un diseño orientado a microservicios, los cuales se comunican con su exterior a través de interfaces RESTful y por medio de mensajes estructurados en formato JSON. Los módulos fueron desplegados a través de contenedores Ubuntu usando Docker Hub.

La capa del controlador cuenta con cuatro módulos que exponen un conjunto de operaciones a ser llamadas desde la vista. Estos módulos o servicios se pueden clasificar en dos categorías: servicios de obtención de información y servicios de procesamiento y análisis de datos. Entre los módulos del prototipo hay uno extra que administra la aplicación llamado functions.py, el cual permite las operaciones CRUD, por las siglas en inglés de create, read, update, delete, sobre los usuarios y el sistema.

3.1.1 SERVICIOS DE OBTENCIÓN DE INFORMACIÓN

El servicio ExtracToJSON procesa la información contenida en metadatos para extraer datos relevantes y facilitar su análisis. El servicio de integrate.py llama a dos procesos ner.py y ner.r (*NER name-entity-recognition*) el cual está basado en displaCy ENT [38], biblioteca que obtiene anotaciones de entidades nombradas formateadas en JSON y que las transforma en HTML semántico. Para la implementación de NER se *tokeniza* y posteriormente se extraen las entidades nombradas. La tokenización divide un texto en segmentos significativos conocidos como *tokens*. La entrada al tokenizer es un texto unicode y el resultado es un objeto con la tokenización correspondiente. El proceso de tokenización, se aplican usando reglas específicas para cada idioma. Adicional a ello, las palabras que son prefijos o sufijos, se eliminan mediante la función stop_word_f, la cual se encarga de quitar a todas aquellas palabras que carecen de un significado por sí mismas.

Para la extracción de entidades nombradas, se crea una instancia que especifica la API y la configuración de spaCy [38]. El lenguaje por defecto en el prototipo es el inglés, pero el prototipo funciona también para el idioma español. Las entidades que se reconocen en el prototipo son: organización, persona, idioma y fecha.

3.1.2 SERVICIOS DE PROCESAMIENTO Y ANÁLISIS DE DATOS

Los análisis básicos, se construyen por medio de histogramas y análisis de frecuencias de diferentes características por ejemplo referencias por años o autores más referenciados.

Para la implementación de clustering se utiliza el lenguaje R [34]. Para el agrupamiento se toman los años de publicación y las palabras clave. En SASR se implementó cluster jerárquico en R [39], al cual le ingresa una matriz X de orden NxP, para la cual, un conjunto de n de documentos, se observa una serie de variables X_1, X_2, ..., X_p. Los datos se estructuran para los n documentos y las p variables que corresponden a los años y las palabras clave. Luego se construyen vectores binarios que representan, para cada documento, si las variables están presentes o no. A partir de X, se construye una matriz S de distancias de orden NxN, donde cada coeficiente S representa la similitud entre dos documentos dada su representación de vectores binarios, así S_{ij} representa el valor del coeficiente de disimilitud para los documentos i y j. Finalmente la matriz de similitud S es la entrada a la función de clustering en R [39].

3.2 CONSTRUCCIÓN DEL CORPUS DE REFERENCIAS

Para la construcción del corpus referencias, se dan unas pautas que se describen a continuación. El primer paso es identificar: palabras clave, autores, revistas, conferencias y centros de investigación relevantes en el tema. Adicional a ello identificar papers de referencia en el tema de investigación. Con la información, que se extrae, de este proceso se construye funciones de búsqueda, las cuales se aplican en sistemas de bases de datos referenciales tales como, WoS (Web of Science), Scopus, Sciencedirect o Scholar Google. Todos estos sistemas permiten almacenar las referencias en un formato común, formato conocido como bibtex. Si se utilizan sistemas de administración de referencias

como Zotero o Mendeley, estos sistemas también permiten exportar la información al formato bibtex o o ris. El resultado inicial de este proceso, se le conoce como lista inicial de referencias. Esta lista de referencias se puede ir actualizando y ajustando en la medida que se conocen nuevos elementos para realizar nuevas funciones de búsqueda, por ejemplo nuevos autores de referencia. Del proceso iterativo de refinamiento de la lista de referencias se obtiene un conjunto de referencias que se conoce como lista filtrada de referencias.

A partir de la lista filtrada de referencias se hace una lectura transversal de los artículos y se revisa los títulos, *abstract* y conclusiones. Esto permite identificar los papers relevantes para el tema y eliminar de la lista de referencias, los artículos que están relacionados con el tema, pero que no son precisamente del tema de investigación. A la lista de referencia se la conoce como lista de referencias del tema enfocado. De igual manera nuevos artículos que son claves para la investigación, se adicionan al conjunto de referencias, la lista de referencias del tema enfocado como su nombre lo indica, es la lista que contiene información de una única área y debe identificar claramente y específicamente el tema particular de investigación. Así, por ejemplo, machine learning no es un tema enfocado, pero sí lo es, tech-mining aplicado servicios IoT en domótica.

Con base en la lista de referencias del tema enfocado se hace una nueva lectura de los documentos con mayor profundidad, en esta lectura se toman anotaciones sobre cada una de las referencias. En sistemas como Zotero o Mendely estas anotaciones se pueden almacenar. La lectura de este proceso debe permitir hacer un nuevo filtro de referencias, así como una nueva inclusión de referencias totalmente relevantes en el tema enfocado, al resultado de este proceso se le conoce como lista de referencias anotada. Como resultado de este proceso se tiene mayor claridad de las preguntas abiertas de investigación. La lista de referencias anotada será el punto de partida para la escritura del estado del arte, el proceso de filtro y actualización de referencias sobre la lista de referencias anotada permitirá obtener el corpus de referencias, ésta es la entrada para el sistema SARS.

Los resultados obtenidos de la herramienta pueden permitir el filtro y actualización de la lista de referencias, así por ejemplo la herramienta permite identificar nuevas palabras clave, palabras clave más específicas sobre el tema o los autores más productivos en el tema. Con esta información se pueden formular ecuaciones de búsqueda, las cuales contienen criterios de inclusión y exclusión de papers, y de esta manera actualizar el corpus de referencias. Este corpus contendría la información sobre los problemas de investigación abiertos, los retos tecnológicos o la cresta de la ola de un tema de investigación particular.

4. RESULTADOS

4.1 MÓDULOS Y FUNCIONALIDAD

El primero módulo permite el registro y login en la aplicación la cual está disponible en [28], la Figura 3 presenta el módulo.

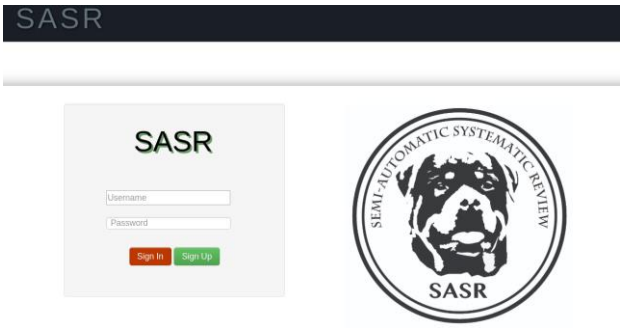

Fig.3. Módulo de registro e inicio de sesión.

Al ingresar al sistema, SASR muestra las siguientes opciones: *Upload File*, *co-authorship, Reports and statistics, Radial Tidy Tree y Treemap*. El sitio principal de SASR muestra la descripción de la herramienta SASR, la descripción y funcionalidad de cada una de las opciones, ver Figura 4.

La opción *Upload File* permite al usuario cargar un archivo de referencias en formato bibtex o ris, el cual contiene el corpus de referencias sobre el tema de investigación. Cuando la carga ha sido finalizada se dirige nuevamente al Sitio principal de SASR y en la parte superior derecha, se muestra el número de referencias que han sido cargadas y procesadas. Las opciones de análisis, se presentan en la siguiente sección.

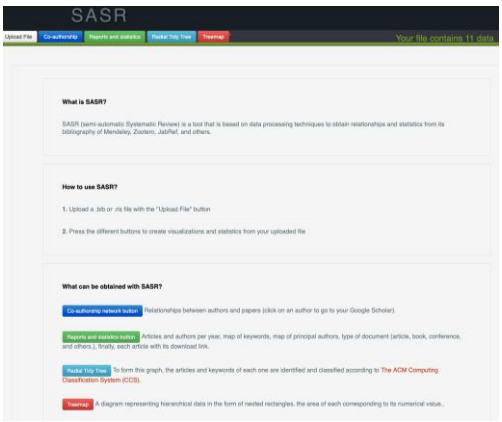

Fig.4. Sitio principal de SASR

4.2 IMPLEMENTACIÓN

La herramienta SASR fue desarrollada usando el lenguaje de programa Python 3 [40], JavaScript [35] y R 3.3.1 [41]. Se utilizaron bibliotecas de clases tales como spaCy [38], D3.js [32] o GoogleCharts.js [36]. La herramienta está alojada en la infraestructura de la Escuela de Ingeniería de Sistemas y Computación de la Universidad del Valle [13].

Semi-Automatic Systematic Review (SASR) es una herramienta que, a partir del corpus de referencias de una temática de investigación,permite ver la dinámica por años de las referencias, ver Figura 5. Los tipos de documentos, los autores más importantes y las palabras clave más relevantes, ver Figura 6 y 7.

La herramienta permite construir la red de coautores y hacer filtros por palabras, año u autores, lo que permite a su vez navegar por el corpus construido, ver Figura 6 y 7.

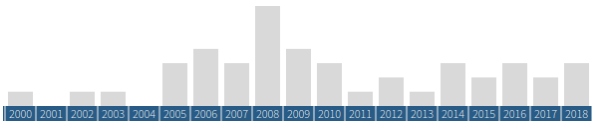

Fig.5. Dinámica de referencias para un tema particular de investigación, computer supported collaborative learning.

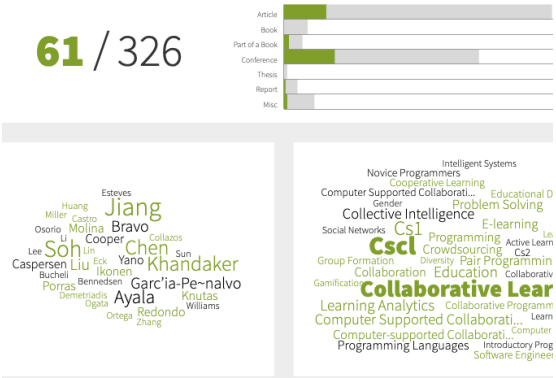

Fig.6. Nube de palabras clave más relevantes para el tema particular *computer supported collaborative learning.*

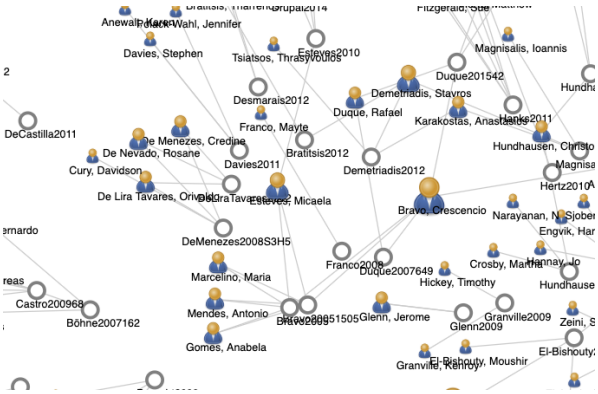

Fig.7. Red de coautoría para un tema particular de investigación, computer supported collaborative learning.

Adicionalmente SASR permite entender la temática en cuestión y utilizar mapas para identificar tendencias, por ejemplo el mapa de áreas ACM (Association for Computing Machinery), ver Figura 8.

Así SASR, se convierte en un software novedoso para entender el estado del arte, permitiendo al investigador navegar y visualizar información con valor agregado acerca de un tema. Por último, se presenta el resultado de análisis de clustering que permite identificar las relaciones entre temas y cuáles de ellos son más relevantes, los colores muestran la información de referencias más actuales, ver Figura 9.

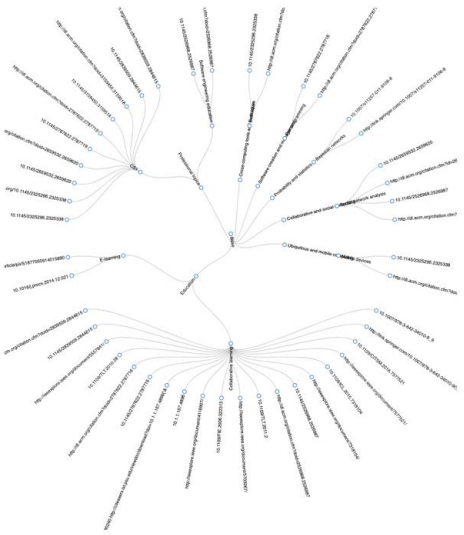

Fig.8. Mapa ACM para un tema particular de investigación, computer supported collaborative learning.

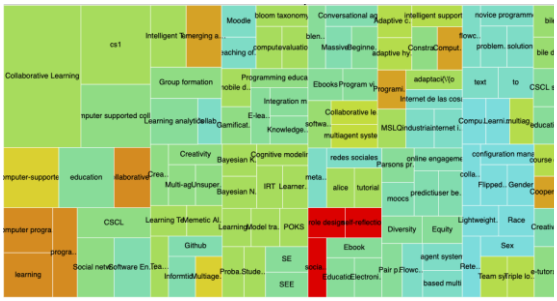

Fig.9. Clustering de temas para computer supported collaborative learning.

4.2.1 ESTUDIO DE CASO

La herramienta se utilizó para la construcción del estado del arte del proyecto de doctorado titulado: un modelo basado en aprendizaje colaborativo asistido por computador y analíticas de aprendizaje para alcanzar competencias en cursos de introducción a la programación. Para el cual se recolectaron 326 referencias: papers, conferencias, entre otros. Las referencias se almacenaron en Mendeley, para ver la lista de referencias [42]. Con base en este corpus de referencias se utilizó la herramienta SASR, la cual permitió encontrar nueva información y clasificaciones que dieron información sobre el tema de investigación.

En el mapa de coautoría, ver Figura 7, se encontró que hay más de 1000 autores relevantes sobre el tema general. Pero en el mapa aparece un grupo de autores de España y Brasil que son los que más han desarrollado el tema CSCL para alcanzar competencias en cursos de programación, con base en esta información se refinó las ecuaciones de búsqueda. A través de las estadísticas de SASR, ver Figuras 5 y 6, se pudo identificar la unión de CSCL e inteligencia artificial, esto tuvo un impacto en la definición del proyecto, desde el título del proyecto mismo. En la línea de tiempo que presenta SASR, se identificó la evolución del uso de tecnologías para el tema de investigación. De igual forma, se pudo identificar que hay 61 artículos que son específicos y relevantes para esta investigación, ver Figura 6.

A través del mapa de la ACM se encontraron referencias no relacionadas con las categorías relacionadas, con el estado de arte que se estaba elaborando; lo cual permitió identificarlas y posteriormente eliminarlas. De igual manera, el mapa de la ACM permitió identificar las áreas en las cuales estaba enmarcada la investigación y las relaciones entre diferentes subtemas que le dan a la investigación el carácter novedoso, ver Figura 8.

Finalmente, a través del mapa de palabras clave, ver Figura 9, se observa que tan relacionadas están las temáticas y cómo unas son interdependientes de otras. En el mapa, se encontraron nuevas claves para realizar nuevas funciones de búsqueda. Así, se toman las palabras en verde (palabras actuales), y se construyen las siguientes funciones de búsqueda en *Web of Science*: TS=(CSCL AND group formation) OR TS=(CS1 AND CSCL AND “cell block method”) OR TS=(CSCL AND “Multi agent” AND POKS). Estos resultados nos llevaron a tener un corpus de referencias más completo que el inicial.

4.3 EVALUACIÓN DE LA HERRAMIENTA SASR

Para evaluar SASR se llevó a cabo una encuesta entre estudiantes de maestría y doctorado que están iniciado proyectos de investigación. Los cuales están en proceso de elaboración de la idea inicial de los trabajos de maestría o doctorado. En total se obtuvieron 40 encuestas, de estudiantes que utilizaron la herramienta. Las cuales permitieron verificar que la herramienta ofrece resultados prometedores. Las respuestas siguen la escala: totalmente en desacuerdo, en desacuerdo, algo en desacuerdo, algo de acuerdo, de acuerdo y totalmente de acuerdo.

En la Figura 10, se muestra que el 79% están de acuerdo con que la herramienta soporta el proceso de escritura de un estado del arte (porcentaje obtenido de sumar de acuerdo y totalmente de acuerdo). Este resultado se confirma con las respuestas a la pregunta ¿Los mapas, gráficos y estadísticas de SASR le dieron información agregada (información que no conocía) sobre su trabajo? a la cual los encuestados responden en igual porcentaje al resultado presentado en la Figura 10.

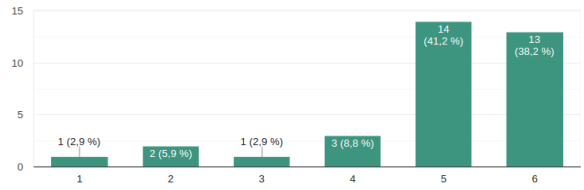

Fig.10. La herramienta SASR ayuda en el análisis de su estado del arte.

Por otra parte, la herramienta permite encontrar nueva información y clasificaciones que dan información sobre el tema de investigación. Esto se verifica en la Figura 11, donde el 32% considera estar en desacuerdo o algo en acuerdo de que SASR permite clasificar las referencias para obtener más información sobre el tema de investigación o principales autores. Por el contrario, el grupo

encuestado restante está en acuerdo de que la herramienta apoya el proceso de obtención de nueva información sobre el tema. Adicionalmente, a la pregunta sobre sí SARS permite encontrar nuevos artículos significativos para su investigación, el 79% de los encuestados respondieron estar de acuerdo.

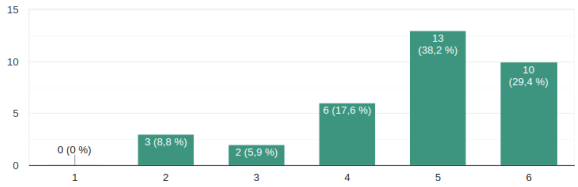

Fig.11. La herramienta permite encontrar nueva información y clasificaciones que dan información sobre el tema de investigación.

Al indagar, si los mapas que presenta la herramienta, permiten obtener información con valor agregado, se verifica con dos preguntas ¿El mapa relacionado con ACM (The ACM Computing Classification System) le permitió entender el área de investigación de su trabajo? y ¿El mapa *treemap* le permitió encontrar relaciones entre temas que desconocía de su investigación? a las cuales más del 70% están de acuerdo con que la herramienta ofrece información con valor agregado, ver Figura 12.

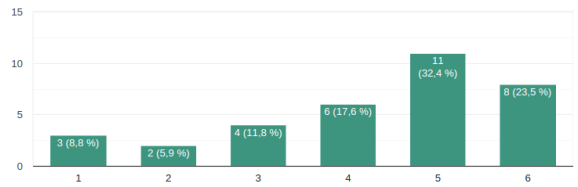

Fig.12. El mapa relacionado con ACM (The ACM Computing Classification System) permite entender el área de investigación.

Finalmente, el 91% reporta que le gusta la herramienta y el 97% la utilizaría en próximas investigaciones.

5. DISCUSIÓN

El desarrollo de la herramienta de software prototipo SASR permitió a los usuarios identificar patrones e información con valor agregado para iniciar un proyecto de desarrollo o investigación. Información tal como principales autores, tipos de publicación, evolución anual, tendencias, palabras claves más relevantes, entre otros.

La herramienta se desarrolló a través de técnicas computacionales para el procesamiento, extracción, análisis y visualización de datos en un modelo cliente servidor. Esto hace que el acceso a la herramienta sea sencillo y de fácil mantenimiento. El modelo vista, datos, controlador y los microservicios hace que el sistema sea flexible y fácilmente extensible. Adicionalmente, su arquitectura cliente servidor permite reducir costos de instalación y de acceso, así como también, permite que la herramienta sea usada por usuarios no expertos.

La Tabla 1. presenta la comparación con herramientas encontradas en la literatura, la cual muestra que hay software pago o que permiten hacer análisis o visualización, pero no todo en uno solo. Otra diferencia es la fuente de información, al utilizar información directamente de las bases de datos referenciales patentes, papers o conferencias, hace que se incurra en costos en tiempo de preprocesamiento de la información, Finalmente ninguna de las herramientas es desarrollada como un servicio desplegado en internet y de libre acceso, lo que hace a la herramienta propuesta, asequible y portable. Adicionalmente, dada la arquitectura de SASR la herramienta es de fácil extensibilidad. Así, otros investigadores pueden incluir nuevos análisis y visualizaciones. La implementación e integración de diversos lenguajes de programación tales como python o R, bibliotecas de clases y algoritmos bajo un modelo robusto, escalable y de bajo costo, permite pensar en el desarrollo en corto tiempo de nuevas funcionalidades.

SASR permite la clasificación y categorización automática de las referencias, los mapas ofrecen información de valor agregado, es decir, las técnicas de inteligencia artificial utilizadas le dan a la herramienta propuesta un carácter diferenciador frente a las otras herramientas ya existentes. Así, las técnicas implementadas de recuperación de información y principalmente de análisis y visualización de datos, permite el estudio de un tema de investigación. Esto disminuye el tiempo y la curva de aprendizaje para analizar el entorno científico en una tema particular, logrando de esta forma entregar información útil para reconocer el estado actual en una investigación.

Tabla 1. Comparación de herramientas existentes y SASR

Software	**Fuente de datos**	**Software libre**	**Análisis / mapas**
Bibexcel	CSV y RIS	Sí	Sí/No
CiteSpace II	CSV	No	No/Sí
CoPalRed	CSV	No	Sí/Sí
VantagePoint	Patentes y documentos	No	Sí/Sí

VOSViewer	CSV	Sí	No/Sí
Vigtech	Scopus	Sí	Sí/Sí
VigHub	Github	Sí	Sí/Sí
SASR	Bibtex	Sí	Sí/Sí

La evaluación de la herramienta por medio de encuestas y un caso de estudio muestra la utilidad de SASR. Principalmente más del 80% de los encuestados reconocen la utilidad de la herramienta y Y manifiestan que la volverían a utilizar, de igual manera encuentran que los resultados arrojados por la herramienta son valiosos para la escritura de un estado del arte.

El prototipo desarrollado es un primer avance para implementar herramientas que permitan conocer el estado del arte de un tema o área específica, allanando así el camino para desarrollar investigaciones en el uso de los técnicas computacionales y prácticas que mejoren los procesos de identificación, extracción, depuración y análisis de información. La herramienta SASR en el contexto local y colombiano puede ser utilidad teniendo en cuenta que no se cuenta con herramientas que den soporte al proceso de formulación de proyectos de investigación y de preguntas de investigación guiadas por una revisión rigurosa de referencias para la escritura del estado del arte.

Como limitaciones del proyecto se encuentra que la herramienta al ser prototipo tiene algunos errores en la lectura de los archivos bibtex, principalmente con caracteres especiales. Otra limitación se refiere a la integración de la herramienta en un curso de introducción a la investigación, la cual mostró nuevos requerimientos u oportunidades de adicionar funcionalidades, por ejemplo, análisis de citaciones. En todo caso la herramienta ha servido como soporte en la construcción de estados del arte en proyectos de maestría y doctorado. Como trabajo futuro se plantea incluir otros formatos de referencias, nuevos análisis y visualizaciones. Finalmente, hacer una evaluación del proceso de construcción del estado del arte asistido por SASR desde una mirada más completa, a través de una metodología cualitativa.

6. AGRADECIMIENTOS

Agradezco a la Universidad del Valle por los recursos y el apoyo para el avance en la investigación. De igual manera Carlos Giovanny Hidalgo por sus conocimientos técnicos, aportes en el tema y su disposición.

7. REFERENCIAS BIBLIOGRÁFICAS

[1] M. Gómez Vargas, C. Galeano Higuita, y D. A. Jaramillo Muñoz, «El estado del arte: una metodología de investigación», *rev.colomb.cienc.soc*, vol. 6, n.º 2, p. 423, jul. 2015.

[2] «Revisiones sistemáticas de la literatura», *Revista Colombiana de Gastroenterologia*, vol. 20, n.º 1, pp. 60-69, mar. 2005.

[3] G. Urrútia y X. Bonfill, «Declaración PRISMA: una propuesta para mejorar la publicación de revisiones sistemáticas y metaanálisis», *Medicina Clínica*, vol. 135, n.º 11, pp. 507-511, oct. 2010.

[4] «state of the art | Definition of state of the art by Lexico», *Lexico Dictionaries | English*. [En línea]. Disponible en: https://www.lexico.com/en/definition/state_of_the_art. [Accedido: 12-sep-2019].

[5] M. Ar, «Language and Ideology in Texts on Globalization: A Critical Discourse Analysis», *IJEL*, vol. 5, n.º 2, p. p63, mar. 2015.

[6] «Zotero | Your personal research assistant». [En línea]. Disponible en: https://www.zotero.org/. [Accedido: 12-sep-2019].

[7] «Mendeley - Reference Management Software & Researcher Network». [En línea]. Disponible en: https://www.mendeley.com/?interaction_required=true. [Accedido: 12-sep-2019].

[8] «BibTeX». [En línea]. Disponible en: http://www.bibtex.org/. [Accedido: 12-sep-2019].

[9] M. J. Cobo, A. G. López-Herrera, E. Herrera-Viedma, y F. Herrera, «Science mapping software tools: Review, analysis, and cooperative study among tools», *J. Am. Soc. Inf. Sci.*, vol. 62, n.º 7, pp. 1382-1402, jul. 2011.

[10] N. J. van Eck y L. Waltman, «Software survey: VOSviewer, a computer program for bibliometric mapping», *Scientometrics*, vol. 84, n.º 2, pp. 523-538, ago. 2010.

[11] B. Johnson y B. Shneiderman, *Tree-maps: A space-filling approach to the visualization of hierarchical information structures*. IEEE, 1991.

[12] J. Stasko y E. Zhang, «Focus+context display and navigation techniques for

enhancing radial, space-filling hierarchy visualizations», en *IEEE Symposium on Information Visualization 2000. INFOVIS 2000. Proceedings*, Salt Lake City, UT, USA, 2000, pp. 57-65.
[13] «SASR». [En línea]. Disponible en: http://eiscapp.univalle.edu.co/uncode/. [Accedido: 17-sep-2019].
[14] G. H. Suarez, *cgiohidalgo/SASR*. 2019.
[15] A. L. Porter y S. W. Cunningham, *Tech mining: exploiting new technologies for competitive advantage*. Hoboken, N.J: Wiley, 2005.
[16] K. Börner, C. Chen, y K. W. Boyack, «Visualizing knowledge domains», *Ann. Rev. Info. Sci. Tech.*, vol. 37, n.º 1, pp. 179-255, ene. 2005.
[17] B. Groneberg-Kloft, D. Quarcoo, y C. Scutaru, «Quality and quantity indices in science: use of visualization tools», *EMBO reports*, vol. 10, n.º 8, pp. 800–803, 2009.
[18] X. Gao y J. Guan, «Networks of scientific journals: An exploration of Chinese patent data», *Scientometrics*, vol. 80, n.º 1, pp. 283–302, 2009.
[19] E. C. Noyons, H. F. Moed, y M. Luwel, «Combining mapping and citation analysis for evaluative bibliometric purposes: A bibliometric study», *Journal of the American society for Information Science*, vol. 50, n.º 2, pp. 115–131, 1999.
[20] «IBM SPSS Software», 05-jun-2019. [En línea]. Disponible en: https://www.ibm.com/co-es/analytics/spss-statistics-software. [Accedido: 12-sep-2019].
[21] «Pajek / PajekXXL / Pajek3XL». [En línea]. Disponible en: http://mrvar.fdv.uni-lj.si/pajek/. [Accedido: 12-sep-2019].
[22] C. Chen, «Searching for intellectual turning points: Progressive knowledge domain visualization», *Proceedings of the National Academy of Sciences*, vol. 101, n.º suppl 1, pp. 5303–5310, 2004.
[23] C. Chen, «CiteSpace II: Detecting and visualizing emerging trends and transient patterns in scientific literature», *J. Am. Soc. Inf. Sci.*, vol. 57, n.º 3, pp. 359-377, feb. 2006.
[24] R. Bailón-Moreno, E. Jurado-Alameda, R. Ruiz-Baños, y J. P. Courtial, «Analysis of the field of physical chemistry of surfactants with the Unified Scientometric Model. Fit», 2005.
[25] R. Bailón-Moreno, E. Jurado-Alameda, y R. Ruiz-Baños, «The scientific network of surfactants: Structural analysis», *Journal of the American Society for Information Science and Technology*, vol. 57, n.º 7, pp. 949–960, 2006.
[26] A. L. Porter y J. Youtie, «How interdisciplinary is nanotechnology?», *J Nanopart Res*, vol. 11, n.º 5, pp. 1023-1041, jul. 2009.
[27] A. L. Porter y J. Youtie, «Where does nanotechnology belong in the map of science?», *Nature Nanotechnology*, vol. 4, p. 534, sep. 2009.
[28] N. J. Van Eck y L. Waltman, «Bibliometric mapping of the computational intelligence field», *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, vol. 15, n.º 05, pp. 625–645, 2007.
[29] N. J. Van Eck, L. Waltman, R. Dekker, y J. van den Berg, «A comparison of two techniques for bibliometric mapping: Multidimensional scaling and VOS», *Journal of the American Society for Information Science and Technology*, vol. 61, n.º 12, pp. 2405–2416, 2010.
[30] F. a. G. Víctor A. Bucheli, «HERRAMIENTA INFORMÁTICA PARA VIGILANCIA TECNOLÓGICA -VIGTECH-», *Avances en Sistemas e Informática*, vol. 4, n.º 1, ene. 2007.
[31] C. G. Hidalgo Suarez, V. A. Bucheli, F. Restrepo-Calle, y F. A. Gonzalez, «A Strategy Based on Technological Maps for the Identification of the State-of-the-Art Techniques in Software Development Projects: Virtual Judge Projects as a Case Study», en *Advances in Computing*, 2018, pp. 338-354.
[32] «The world's leading software development platform · GitHub». [En línea]. Disponible en: https://github.com/. [Accedido: 18-sep-2019].
[33] M. Pérez-Herrera Cuadrillero, «Arquitecturas basadas en microservicios», 2015. [En línea]. Disponible en: http://oa.upm.es/37346/. [Accedido: 17-sep-2019].
[34] «Flask», *Pallets*. [En línea]. Disponible en: https://palletsprojects.com/p/flask/.

[Accedido: 17-sep-2019].
[35] «Free JavaScript training, resources and examples for the community». [En línea]. Disponible en: https://www.javascript.com/. [Accedido: 17-sep-2019].
[36] «Charts», *Google Developers*. [En línea]. Disponible en: https://developers.google.com/chart/. [Accedido: 17-sep-2019].
[37] M. Bostock, «D3.js - Data-Driven Documents». [En línea]. Disponible en: https://d3js.org/. [Accedido: 17-sep-2019].
[38] «spaCy · Industrial-strength Natural Language Processing in Python». [En línea]. Disponible en: https://spacy.io/. [Accedido: 17-sep-2019].
[39] «Hierarchical cluster analysis on famous data sets - enhanced with the dendextend package». [En línea]. Disponible en: https://cran.r-project.org/web/packages/dendextend/vignettes/Cluster_Analysis.html. [Accedido: 17-sep-2019].
[40] «Python 3.0 Release», *Python.org*. [En línea]. Disponible en: https://www.python.org/download/releases/3.0/. [Accedido: 17-sep-2019].
[41] «The Comprehensive R Archive Network». [En línea]. Disponible en: https://cran.r-project.org/. [Accedido: 17-sep-2019].
[42] «References CSCL approach for CS1. | Mendeley». [En línea]. Disponible en: https://www.mendeley.com/community/state-of-the-art-17/. [Accedido: 18-sep-2019].