Big Data Optimization: Framework Algorítmico para el análisis de Datos guiado por Semántica

Cristóbal Barba González
Director: José Francisco Aldana Montes
Director: José Manuel Gracía Nieto
Grupo de investigación Khaos
Edificio de investigación Ada Byron
Departamento de Lenguajes y Ciencias de la Computación
Universidad de Málaga, España.
cbarba@lcc.uma.es
Inicio: Marzo 2015

***Resumen*—En las últimas décadas el aumento de fuentes de información en diferentes campos de la sociedad desde la salud hasta las redes sociales, ha puesto de manifiesto la necesidad de nuevas técnicas para su análisis, lo que se ha venido a llamar el *Big Data*. Los problemas clásicos de optimización no son ajenos a este cambio de paradigma, como por ejemplo el problema del viajante de comercio (TSP), ya que se puede beneficiar de los datos que proporciona los diferentes sensores que se encuentran en las ciudades y que podemos acceder a ellos gracias a los portales de *Open Data*. En esta tesis se ha desarrollado un nuevo framework, jMetalSP, para la optimización de problemas en el ámbito del Big Data permitiendo el uso de fuentes de datos externas para modificar los datos del problema en tiempo real. Por otro lado, cuando estamos realizando análisis, ya sea de optimización o machine learning en Big Data, una de las formas más usada de abordarlo es mediante workflows de análisis. Estos están formados por componentes que hacen cada paso del análisis. El flujo de información en workflows puede ser anotada y almacenada usando herramientas de la Web Semántica para facilitar la reutilización de dichos componentes o incluso el workflow completo en futuros análisis, facilitando así, su reutilización y a su vez, mejorando el procesos de creación de los mismos. Para ello se ha creado la ontología *BIGOWL*, que permite trazar la cadena de valor de los datos de los workflows mediante semántica y además ayuda al analista en la creación de workflow gracias a que va guiando su composición con la información que contiene por la anotación de algoritmos, datos, componentes y workflows.**

***Index Terms*—Big Data, Optimización, Machine Learning, Web Semantic**

I. Introducción

Debido al gran avance que existe día a día en las tecnologías de información, las organizaciones se han tenido que enfrentar a nuevos desafíos que les permitan analizar, descubrir y entender más allá de lo que sus herramientas tradicionales reportan sobre su información. Existen numerosas fuentes de información como son, las aplicaciones disponibles en internet (redes sociales, geo-referenciamiento, etc.); la información obtenida en el campo de la medicina y biología con la secuenciación del ADN y otros datos analíticos, etc. Todos estos datos se conocen como *Big Data* y hace necesario el uso de nuevos enfoques para el análisis de los mismos, debido principalmente, a que las técnicas tradicionales usadas hasta ahora (Machine Learning y Optimización) no están diseñadas ni optimizadas para manejar grandes volúmenes de información [1].

El término Big Data por tanto, hace referencia a aquellos datos que no pueden ser procesados o analizados usando las técnicas tradicionales [2]. El análisis del Big Data permite extraer información a partir de estos datos. Multitud de soluciones software alrededor del proyecto Apache Hadoop van resolviendo las diferentes problemáticas a través de este proyecto y otros complementarios. Alguno de estos proyectos son:

- ***Apache Hadoop***. Que integra el framework MapReduce y el sistema de archivos HDFS.
- ***Apache Spark***. Framework que permite realizar tareas de transformación y de acción sobre los datos en streaming de manera eficiente.

De acuerdo con Gartner [1] y la asociación europea Big Data Value (BDVA)[2], hay un reto en el campo del Big Data acerca de la construcción de aplicaciones que inyecte el conocimiento del dominio (problema, algoritmo, dato, etc.), así como el contexto, en el proceso de análisis. Por contexto entendemos toda la (meta)información relevante que permita interpretar los resultados del análisis. Esto tendrá como consecuencia la accionabilidad de dichos resultados. Así facilitará la interpretación de estos datos; permitirá integrarlos fácilmente con otros datos estructurados; facilitará la integración del sistema de análisis del Big Data con otros sistemas y habilitará la interconexión de algoritmos. En esta tesis hemos abordado este reto creando BIGOWL, una ontología en la que se ha definido toda la semántica necesaria para poder definir cualquier problema de análisis en el ámbito del Big Data así como los algoritmos y datos a utilizar para dicho análisis.

Uno de los principales problemas que nos encontramos a la hora de trabajar con Big Data es la eficiencia de los algoritmos

[1]https://www.gartner.com/doc/3656517/adopt-datadriven-approach-consolidating-infrastructure

[2]http://www.bdva.eu/

de optimización a la hora de procesar grandes volúmenes de información, sobre todo en tiempo de cómputo, por ello es necesario realizar algoritmos optimizados para problemas de Big Data. Existen dos grandes tipos de problemas de optimización los llamados mono-objetivo y los multi-objetivo. Este trabajo se ha centrado en los problemas multi-objectivo ya que existen muchos en Big Data. Para ello hemos desarrollado *jMetalSP* que es una librería de algoritmos de optimización adaptados para Big Data, el cual está basado en jMetal [3], una herramienta para la optimización multi-ojetivo y Apache Spark [4]. Dentro de los algoritmos multi-objetivo desarrollados en jMetalSP, nos hemos centrado en los algoritmos dinámicos y en los interactivos, ya que facilitan reducir el espacio de búsqueda de soluciones de un problema y por tanto, reducir el tiempo de cómputo quedándose sola con las áreas preferidas por el analista (Decision Maker). Además permiten abordar procesamiento en streaming (Velocidad), actuar ante cambios en el problema, datos o entorno (Variabilidad) y acercar el proceso de optimización al Decision Maker (Valor, Veracidad).

II. Antecedente e Hipótesis de Partida

Los problemas de procesamiento que dan lugar al Big Data provienen de la cantidad de datos, que es desproporcionadamente grande y del origen de la información, que es muy variado (fuentes, formato, etc.). Estos problemas asociados al Big Data se entienden en cinco dimensiones (las denominadas "cinco uves"): Volumen, Velocidad, Variedad, Variabilidad y Veracidad. El Big Data trae consigo la posibilidad de encontrar información relevante en los nuevos tipos de datos emergentes la capacidad de abordar cuestiones que hasta no hace mucho eran imposibles de plantear. El análisis del Big Data puede revelar información importante que anteriormente permanecía oculta debido al elevado coste de procesar ciertos datos, tales como las tendencias sociales de opinión, tendencias de consumo de ciertos productos comerciales, orientación y seguimiento publicitario, etc.

En general, la generación de un análisis en Big Data es compleja ya que entran en juego diferentes aplicaciones o algoritmos, concretamente diferentes componentes, que se tienen que interconectar entre ellos y tienen que ser compatibles, es decir, los datos de salida de uno son los de entrada del siguiente. En este sentido, esta tesis se propone que el uso de la semántica apoyado con tecnologías de la Web Semántica puede ayudar en la anotación de toda la información necesaria para que componentes distintos puedan interactuar entre ellos, facilitando así la interconexión entre componentes, su reutilización y la trazabilidad de la cadena de valor de los datos.

La problemática se encuentra en dar estructura a esta información para poder ser integrada en los componentes. De hecho, la automatización de este proceso es uno de los grandes retos en los que las tecnologías relacionadas con los estándares de la Web Semántica podrían tener un papel relevante.

Hoy en día se están desarrollando una serie de plataformas para el análisis del Big Data que además ofrecen una conexión con otros sistemas para la asistencia en la toma de decisiones. No obstante, hasta donde llega nuestro conocimiento, estos frameworks no incluyen una anotación semántica de sus elementos (componentes, algoritmos, métodos, interfaces, etc.) que facilite la reutilización y la composición de workflows de análisis del Big Data. Tampoco aprovechan la semántica específica del dominio de aplicación de forma que se facilite el desarrollo de los workflows de análisis, la anotación de los datos y los metadatos y/o de los resultados de los algoritmos, además de la generación de nuevos operadores adaptados que utilizan la estructura semántica del dominio del problema. Esto provoca que:

1. Los algoritmos de análisis del Big Data no son fácilmente reutilizables, no son capaces de explotar la semántica y, por tanto, la creación de workflows "inteligentes" es complicado y requiere desarrollo ad hoc.
2. No existe un procedimiento estandarizado para conectar los resultados del análisis del Big Data con otros sistemas.

Todas estas razones nos llevan a plantear la principal hipótesis de este proyecto: *la "anotación semántica"de los componentes software que constituyen un framework para el análisis del Big Data puede actuar como nexo de unión, tanto para la generación de nuevas propuestas algorítmicas que utilicen este conocimiento semántico, como para su final conexión con otros sistemas*. Se reutiliza así la idea fundamental de la Web Semántica en el contexto del Big Data orientado a la optimización.

Por otro lado, hemos desarrollado un framework para el análisis del Big Data, centrándonos en la optimización multi-objectivo para Big Data, más concretamente en metaheurísticas dinámicas y/o interactivas. Este framework, junto con otros de Machine Learning (WEKA, Scikit Learn Python, MLlib Spark, etc.), se usarán como casos de uso para comprobar que la anotación de sus componentes es capaz de cubrir completamente sus especificaciones.

Los algoritmos que se han desarrollado en el framework se usarán en diferentes problemas reales, como son: la bioinformática o la búsqueda de la ruta óptima en tráfico. Hemos integrado el framework jMetal, como base, que proporciona una gran batería de algoritmos de optimización multi-objetivo, junto con el modelo de programación MapReduce [5] y el framework Spark [4].

Otro de nuestros objetivos es realizar un benchmark para los algoritmos interactivos que se desarrollen en esta tesis y que nos permita compararlos ya que solo existe uno en el estado de arte actual [6].

III. Objetivos

Esta tesis tiene como objetivo el integrar técnicas y resultados del análisis del Big Data con una capa de metadatos (de los datos objetos del análisis, de las técnicas de análisis y del dominio donde éstas se aplican) para romper las barreras de acceso y aplicabilidad relacionados con las tecnologías de análisis del Big Data. Como enfoque principal el desarrollo de herramientas para dicho análisis, nos centramos en la optimización en Big Data. Concretamente los principales objetivos son:

O 1. Definir un modelo ontológico para la anotación semántica de algoritmos de análisis del Big Data.
  - O 1.1. Desarrollar una clasificación semántica de los algoritmos, componentes de procesamiento y visualización.
  - O 2.2. Diseñar una metodología para anotar la funcionalidad genérica de los algoritmos (tipo de algoritmos, entradas, salidas, transformación de los datos).
  - O 3.3. Diseñar una metodología para anotar las entradas, salidas y tipo de algoritmos según una ontología de dominio en caso de ser algoritmos específicos de un dominio de aplicación.

O 2. Desarrollo de una plataforma para la optimización en problemas Big Data.
  - O 2.1. Estudio de nuevos algoritmos de análisis en Big Data (dinámicos y/o interactivos).
  - O 2.2. Desarrollo de mecanismos para la evaluación de algoritmos interactivos.
  - O 2.3. Diseñar la estructura del repositorio para incluir no sólo los algoritmos, sino también sus anotaciones semánticas relativas a su funcionalidad y los eventos que generan o consumen.

O 3. Validación de la plataforma con casos de uso reales y académicos.
  - O 3.1. Problema del viajante de comercio en la ciudad de Nueva York con streaming Open Data.
  - O 3.2. Inferencia en Redes génicas (E.coli, Yeast).
  - O 3.3. Familia de problemas multi-objectivo DTLZ.
  - O 3.4. Familia de problemas dinámicos y multi-objetivo FDA.

IV. Metodología y Plan de Trabajo

El método a utilizar es una adaptación del método Investigación en Acción de Avison, et al. [7]. Se trata de un método cualitativo utilizado para validar los trabajos de investigación mediante su aplicación a proyectos reales. En la Conferencia sobre Procesamiento de Información de 1998 se declaró a los métodos cualitativos como métodos de investigación apropiados para el campo de los sistemas de información [7], y los difundió en [8]. Proponemos un método de investigación genérico, basado en la propuesta de Bunge [9] y formado por etapas que, dada su generalidad, pueden aplicarse a cualquier tipo de investigación. Por tanto, utilizaremos un método de validación práctica, especialmente apropiado para la investigación en ingeniería y específicamente para validar aquellos resultados, en los que la aplicación de un método científico tradicional (inductivo o hipotético-deductivo) no son directamente aplicables. El método a seguir para la resolución y validación del problema concreto que nos ocupa (método de Investigación en Acción) no es un proceso lineal, sino que va avanzando mediante la compleción de ciclos. Al comenzar cada ciclo se ponen en marcha nuevas ideas, que son puestas en práctica y comprobadas hasta el inicio del siguiente ciclo [10]. Este proceso cíclico, en el que hemos ido probando y refinando cada uno de los resultados obtenidos, será nuestro método de validación.

La metodología Software que se va a seguir estará inspirada en este método, ya que se tendrán como referencia modelos ágiles como Scrum [3], basados en iteraciones de breve duración.

En concreto las tareas realizadas, para cada ciclo, son:

- ***Observación***. Estudio pormenorizado del problema a tratar, realizando un estudio del estado del arte e identificando posibles riesgos antes de afrontar la tarea.
- ***Formulación de hipótesis***. Declaración de la hipótesis que queremos llevar a cabo en dicho ciclo, se dividirá en pequeñas tareas abordables en la duración del ciclo.
- ***Recogida de observaciones***. Obtención de resultados como consecuencia de la realización de las tareas del paso anterior.
- ***Contrastes de hipótesis***. Estudio de las observaciones recogidas y comprobación si se han cumplido nuestras hipótesis de partida.
- ***Demostración o refutación de hipótesis***. Aceptación o rechazo de la hipótesis, si es necesario una modificación de la misma, empieza un nuevo ciclo con estos cambios.

Para alcanzar los objetivos siguiendo esta metodología, se definieron los siguientes pasos del plan de trabajo:

1. Análisis de la tarea, estudiando el problema que queremos abordar, tecnología a usar y búsqueda de otras propuestas en el estado del arte.
2. Diseño de la tarea, estudio de detalles técnicos y patrones de diseño a usar, ya sea para aplicarlo a la implementación de nuevos algoritmos o la anotación de componentes de BIGOWL (ontología propuesta).
3. Implementación del diseño anteriormente indicado.
4. Validación de los resultados obtenidos mediante el uso de herramientas externas como pueden ser estadísticos, razonadores de ontología, etc.
5. Difusión y publicación de revistas y congresos internacionales de impacto o relevancia científica.

V. Relevancia de la tesis

Esta tesis está financiada mediante la beca *BES-2015-072209* del proyecto PERCEPTION [TIN 2014-58304-R] del Ministerio de Economía y Competitividad de España, además, se alinea directamente con las prioridades del PLAN ESTATAL DE INVESTIGACIÓN CIENTÍFICA Y TÉCNICA Y DE INNOVACIÓN 2013-2016 al plantear la generación de conocimiento en el ámbito de Big Data, y ser parte de la estrategia del grupo de investigación al que pertenezco, *grupo de investigación Khaos*[4] [5]. Por otro lado, la línea que se está siguiendo en dicha tesis, está en concordancia con las áreas prioritarias del Octavo Programa Marco de la Unión Europea Horizonte 2020, ICT–19- [2018-20]: Big Data PPP: *Methods and tools for extreme-scale analytics, and innovation hubs* en

[3] https://www.scrum.org
[4] http://khaos.uma.es/
[5] http://khaos.uma.es/cbarba

el área de "investigación e innovación", en la que se esperan proyectos para desarrollar entre otros, algoritmos, arquitecturas software y metodologías de optimización novedosas para el análisis del Big Data.

En la Figura 1. se pueden ver las principales líneas de investigación que se están siguiendo en esta tesis. Como elemento central tenemos la optimización en Big Data la cual se apoya en los otros conceptos, como por ejemplo, los modelos de la web semántica para la anotación de los componentes de los workflows con el fin de mejorar el proceso de creación de los mismos, o la optimización multi-objetivo junto con los procedimientos en streaming para crear la optimización multi-objetivo dinámica con el fin de poder abordar optimizaciones con grandes volúmenes de datos.

Con el objetivo de poner en relieve la relevancia de esta tesis, se muestra a continuación la lista de las publicaciones realizadas durante su desarrollo:

1. ***Fine Grain Sentiment Analysis with Semantics in Tweets*** [11]. Este artículo se realizó una primera aproximación al mundo del Big Data añadiéndole semántica al análisis de los datos. Se realizó un estudio de los sentimientos de los aficionados durante el torneo universitario de baloncesto de Estados Unidos *Big 12 Men's Basketball Championship*. En este artículo se comprobó de manera empírica, nuestra hipótesis, la semántica mejora y facilita el análisis en Big Data.
2. ***Dynamic Multi-Objective Optimization With jMetal and Spark: a Case Study*** [12]. El objetivo de este artículo fue comprobar que el uso de un framework de Big Data como Spark facilita la lectura de datos en streaming para su optimización. Es decir, se comprobó que era posible dotar con datos reales y en streaming a un problema de optimización.
3. ***Un Framework para Big Data Optimization Basado en jMetal y Spark*** [13]. En este trabajo se presenta una primera aproximación al framework para la optimización en Big Data, en el que se utiliza Spark como motor de computación distribuida y jMetal como motor de optimización.
4. ***jMetalSP: a framework for dynamic multi-objective big data optimization*** [14]. En este artículo se presenta la primera versión de jMetalSP, la herramienta para la optimización de Big Data. En esta primera versión jMetalSP está compuesto por una serie de algoritmos de optimización multi-objectivo y dinámicos. Se usa como caso de uso el problema de TSP con datos reales de la ciudad de Nueva York en streaming.
5. ***Multi-Objective Big Data Optimization with jMetal and Spark*** [15]. Con este trabajo se evalúa el rendimiento de jMetalSP y la plataforma de análisis con 100 nodos que hemos configurado para ejecutar todo el software desarrollado en esta tesis. Se comprueba la escalabilidad de la plataforma y del propio framework. Para ello se evalúan los tiempos de computación tratando con diferentes tamaños de datos y complejidad algorítmica del problema. Así como el número de nodos usados. Finalmente, una comparativa entre los framework MapReduce de Hadoop y Spark desde un punto de vista de tiempo de ejecución en algoritmos de optimización.
6. ***Design and Architecture of the jMetalSP Framework*** [16]. Este trabajo tiene como objetivo la presentación de una nueva arquitectura para jMetalSP en el que, ya no es tan dependiente de Spark, sino que se hace flexible a cualquier motor de ejecución paralela. Se añaden nuevos algoritmos dinámicos como Dynamic SMPSO y Dynamic MOCell, así como la familia de problemas dinámicos FDA para su evaluación.
7. ***InDM2: Interactive Dynamic Multi-Objective Decision Making Using Evolutionary Algorithms*** [17]. En este artículo se presenta el primer algoritmo interactivo y dinámico de optimización multi-objetivo al que hemos llamado InDM2. El desarrollo de esta nueva propuesta algorítmica se ha realizado usando jMetalSP, por lo que añadimos a la plataforma algoritmos interactivos además de dinámicos.
8. ***BIGOWL: Knowledge Centered Big Data Analytics*** [18]. En este artículo se presenta BIGOWL, la propuesta ontológica para el análisis del Big Data. Define todos los componentes necesarios para el diseño, construcción y ejecución de un workflow. Así como los elementos de más bajo nivel que los conforman (algoritmos, problemas, datos, tareas,etc). Además se presentan una serie de consultas SPARQL y reglas de razonamiento para guiar el proceso de creación y validación de workflows. Se utiliza jMetalSP y el caso de uso del problema TSP con datos reales de la ciudad de Nueva York para validar el modelo semántico.
9. ***Análisis de datos de acelerometría para la detección de tipos de actividades*** [19]. Se realiza un estudio de viabilidad para la clasificación de actividades físicas obtenidas mediante pulseras de acelerometría en pacientes con problemas cardiovasculares, utilizando para ello algoritmos de *Deep Convolutional Neural Networks* (una modalidad de ConvNet) para un conjunto de datos de más de 4 TB.
10. ***Artificial Decision Maker Driven by PSO: An Approach for Testing Reference Point Based Interactive Methods*** [20]. Este trabajo presenta una seria de mecanismos para la evaluación de algoritmos interactivos de forma automática. Con esta aproximación es posible evaluar cualquier algoritmo interactivo que utilice puntos de referencias para indicar las preferencias del usuario. Esta propuesta se basa en la búsqueda del algoritmo PSO para encontrar buenas soluciones óptimas o cercanas a ellas. En nuestro caso, usamos la búsqueda de PSO para indicarle al algoritmo interactivo un punto de referencia cercano al punto de aspiración facilitado por el analista o Decision Maker.
11. ***Extending the Speed-constrained Multi-Objective PSO (SMPSO) With Reference Point Based Preference Articulation*** [21]. En este artículo presentamos una nueva versión interactiva del algoritmo SMPSO, usando el fra-

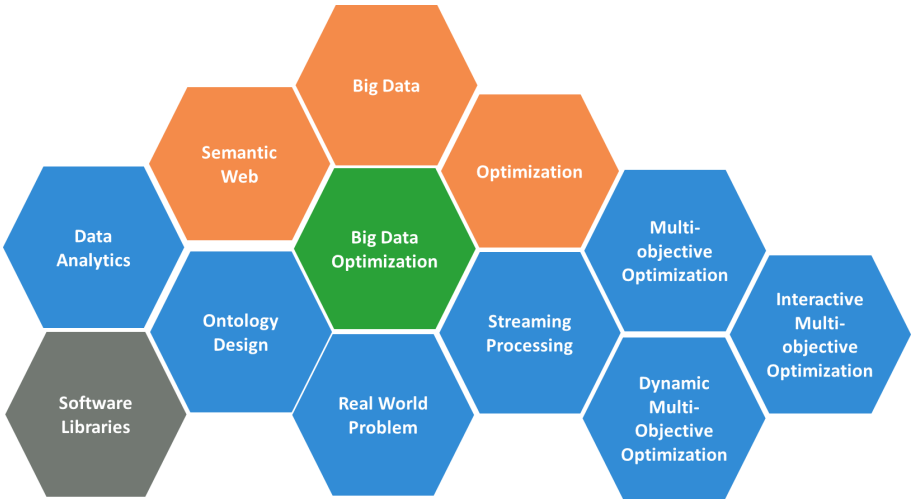

Figura 1. Líneas de investigación en la tesis.

mework jMetalSP. SMPSO con puntos de referencia puede centrarse solo en un área del espacio de búsqueda del problema especificado por el analista mediante los puntos de referencias, mejorando así los tiempos de ejecución ya que se reduce mucho el espacio de búsqueda.

12. ***Scalable Inference of Gene Regulatory Networks with the Spark Distributed Computing Platform*** [22]. En este trabajo resolvemos el problema biológico de la inferencia en redes génicas usando el jMetalSP ya que, gracias a Spark, se puede realizar computación distribuida y permite resolver este problema disminuyendo el tiempo de ejecución.
13. ***About Designing an Observer Pattern-Based Architecture for a Multi-ObjectiveMetaheuristic Optimization Framework*** [23]. Este trabajo versa sobre la presentación de una nueva arquitectura para jMetal usando el patrón observador en el que se podrá componer metaheurísticas mediante componentes independientes y reusables. Por tanto, extendemos la idea de la composición de workflow a la composición de algoritmos.

En la Figura 2. se muestra un resumen de los diferentes trabajos antes mencionados junto con los tópicos de investigación que se cubren en esta tesis.

VI. Conclusiones

Las concusiones que podemos destacar en esta tesis son:

- Esta tesis pretende abordar una serie de problemas actuales alineados con las prioridades estratégicas de los planes de investigación regional, nacional y europea.
- Se ha generado software, como son jMetalSP y BIGOWL, que aportan valor a la comunidad científica y empresarial.
- Presenta una serie de contribuciones científicas bien avaladas en términos de publicaciones

Como futuras líneas de trabajo se pretende desarrollar nuevas propuestas algorítmicas que hagan uso del conocimiento semántico que se pueda extraer del domino de los problemas, facilitando así sus resoluciones.

Referencias

[1] M. Chen, S. Mao, and Y. Liu, "Big data: A survey," *Mobile networks and applications*, vol. 19, no. 2, pp. 171–209, 2014.
[2] G.-H. Kim, S. Trimi, and J.-H. Chung, "Big-data applications in the government sector," *Communications of the ACM*, vol. 57, no. 3, pp. 78–85, 2014.
[3] J. J. Durillo and A. J. Nebro, "jmetal: A java framework for multi-objective optimization," *Advances in Engineering Software*, vol. 42, no. 10, pp. 760–771, 2011.
[4] M. Zaharia, M. Chowdhury, M. J. Franklin, S. Shenker, and I. Stoica, "Spark: Cluster computing with working sets," in *Proceedings of the 2Nd USENIX Conference on Hot Topics in Cloud Computing*, ser. HotCloud'10. USENIX Association, 2010, pp. 10–10.
[5] K. Shvachko, H. Kuang, S. Radia, and R. Chansler, "The hadoop distributed file system," in *Mass storage systems and technologies (MSST), 2010 IEEE 26th symposium on*. Ieee, 2010, pp. 1–10.
[6] V. Ojalehto, D. Podkopaev, and K. Miettinen, "Towards automatic testing of reference point based interactive methods," in *International Conference on Parallel Problem Solving from Nature*. Springer, 2016, pp. 483–492.
[7] D. E. Avison, F. Lau, M. D. Myers, and P. A. Nielsen, "Action research," *Communications of the ACM*, vol. 42, no. 1, pp. 94–97, 1999.
[8] C. B. Seaman, "Qualitative methods in empirical studies of software engineering," *IEEE Transactions on software engineering*, vol. 25, no. 4, pp. 557–572, 1999.
[9] M. Bunge, "La investigación científica," *Ariel S.A.*, 1976.
[10] P. Reason and H. Bradbury, *Handbook of action research: Participative inquiry and practice*. Sage, 2001.
[11] C. Barba-González, J. García-Nieto, I. N. Delgado, and J. F. A. Montes, "A fine grain sentiment analysis with semantics in tweets," *IJIMAI*, vol. 3, no. 6, pp. 22–28, 2016.
[12] J. A. Cordero, A. J. Nebro, C. Barba-González, J. J. Durillo, J. García-Nieto, I. Navas-Delgado, and J. F. Aldana-Montes, "Dynamic multi-objective optimization with jmetal and spark: a case study," in *LNCS of International Workshop on Machine Learning, Optimization and Big Data (MOD 2016)*. Springer, 2016, pp. 106–117.
[13] C. Barba-González, A. J. Nebro, J. García-Nieto, J. A. Cordero, J. J. Durillo, I. Navas-Delgado, and J. F. Aldana-Montes, "Un framework para big data optimization basado en jmetal y spark," in *LNCS of XI Congreso Español de Metaheurísticas, Algoritmos Evolutivos y Bioinspirados (MAEB 2016)*, 2016, pp. 159–168.
[14] C. Barba-González, J. García-Nieto, A. J. Nebro, J. A. Cordero, J. J. Durillo, I. Navas-Delgado, and J. F. Aldana-Montes, "jmetalsp: a framework for dynamic multi-objective big data optimization," *Applied Soft Computing*, vol. 69, pp. 737–748, 2017.

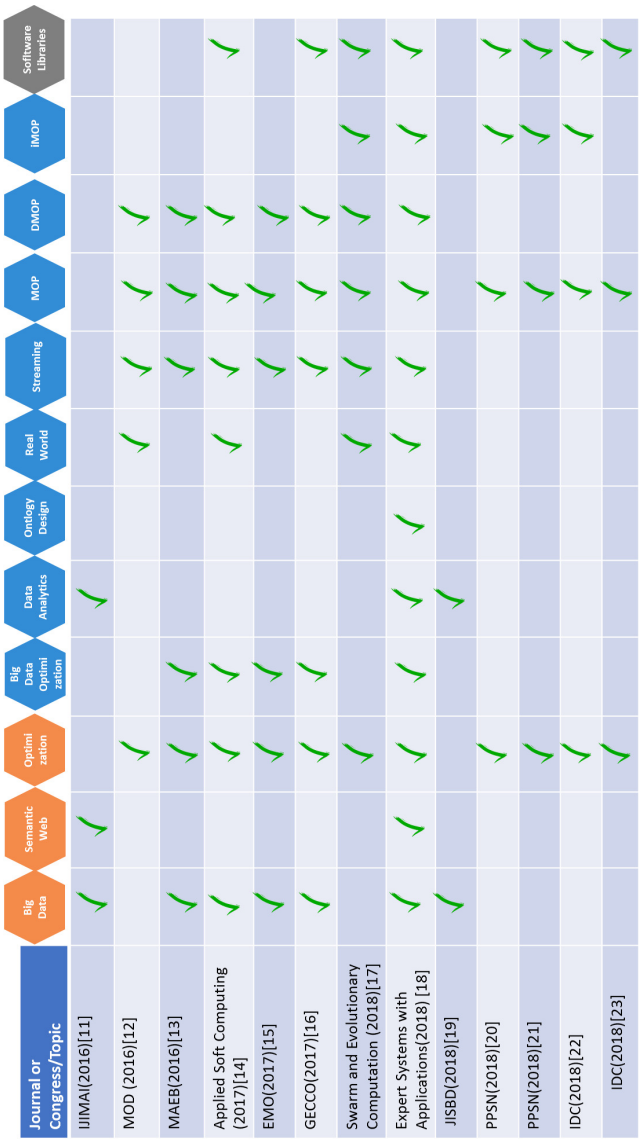

Figura 2. Research contributions in this thesis.

[15] C. Barba-Gonzaléz, J. García-Nieto, A. J. Nebro, and J. F. Aldana-Montes, "Multi-objective big data optimization with jmetal and spark," in *LNCS of International Conference on Evolutionary Multi-Criterion Optimization (EMO'17), GGS class 2 (CORE A)*. Springer, 2017, pp. 16–30.

[16] A. J. Nebro, C. Barba-González, J. García-Nieto, J. A. Cordero, and J. F. A. Montes, "Design and architecture of the jmetaisp framework," in *Proceedings of the Genetic and Evolutionary Computation Conference Companion (GECCO'17)*. ACM, 2017, pp. 1239–1246.

[17] A. J. Nebro, A. B. Ruiz, C. Barba-González, J. García-Nieto, M. Luque, and J. F. Aldana-Montes, "Indm2: Interactive dynamic multi-objective decision making using evolutionary algorithms," *Swarm and Evolutionary Computation*, vol. 40, pp. 184–195, 2018.

[18] C. Barba-González, J. García-Nieto, M. d. M. Rodan-García, I. Navas-Delagado, and J. F. Aldana-Montes, "Bigowl: Knowledge centered big data analytics," *Expert Systems with Applications*, vol. 115,, pp. 543–556, 2018.

[19] S. Huratdo-Requena, C. Barba-González, M. Rybiński, F. J. Barón-López, J. Wärnberg, I. Navas-Delgado, and J. F. Aldana-Montes, "Análisis de datos de acelerometría para la detección de tipos de actividades," in *Jornadas de Ingeniería del Software y Bases de Datos (In press)*, 2018.

[20] C. Barba-González, V. Ojalehto, J. García-Nieto, A. J. Nebro, K. Miettinen, and J. F. Aldana-Montes, "Artificial decision maker driven by pso: An approach for testing reference point based interactive methods," in *Proceeding of 15th International conference on parallel problem solving from nature (PPSN'18), GGS A class 2(CORE A)*. Springer, 2018, pp. 274–285.

[21] A. J. Nebro, J. J. Durillo, J. García-Nieto, C. Barba-González, J. Del Ser, C. A. Coello Coello, A. Benítez-Hidalgo, and J. F. Aldana-Montes, "Extending the speed-constrained multi-objective pso (smpso) with reference point based preference articulation," in *Proceeding of 15th International conference on parallel problem solving from nature (PPSN'18), GGS A-, CORE A*. Springer, 2018, pp. 298–310.

[22] C. Barba-González, J. García-Nieto, A. J. Nebro, A. Benítez-Hidalgo, and J. F. Aldana-Montes, "Scalable inference of gene regulatory networks with the spark distributed computing platform," in *Springer Series of 12th International Symposium on Intelligent Distributed Computing (IDC'18)*, 2018.

[23] A. Benítez-Hidalgo, A. J. Nebro, J. J. Durilllo, J. García-Nieto, E. Camacho-López, C. Barba-González, B, and J. F. Aldana-Montes, "About designing an observer pattern-based architecture for a multi-objectivemetaheuristic optimization framework," in *12th International Symposium on Intelligent Distributed Computing*, 2018.