

الجمهورية الشعبية الديمقراطية الجزائرية
République Algérienne Démocratique et Populaire
وزارة التعليم العالي والبحث العلمي
Ministère de l'Enseignement Supérieur et de la Recherche Scientifique
المدرسة العليا للإعلام الآلي 080 ماي 5491. بسيدي بلعباس
École Supérieure en Informatique
-08 Mai 1945- Sidi Bel Abbès



THESIS

To obtain the diploma of **Engineer**
Field: **Computer Science**
Specialty: **Intelligence Artificielle et Science des Données (IASD)**

Theme

**Multi-Agent System for Automating Web Scraping: A
Scalable Framework for Domain-Specific RAG-Based
Question Answering**

Presented by:
HAGANI ABLA

Submission Date: **Sept 30th, 2025**
In front of the jury composed of:

Dr. KLOUCHE BADIA
Dr. BOUSMAHA Rabab
Pr. BENSLIMANE Sidi Mohamed
Pr. ZIOU Djemel

President
Examiner
Supervisor
Co-Supervisor

Academic Year : 2024/2025

Abstract

Traditional web scraping methods face significant scalability and maintenance challenges when dealing with modern JavaScript-heavy websites, anti-bot mechanisms, and dynamic content. This thesis proposes a multi-agent system that automates large-scale web scraping and knowledge extraction by distributing tasks across specialized agents responsible for URL crawling, content rendering, data parsing, and semantic indexing.

The proposed framework ensures scalability through parallel task execution, robust error handling, and modular maintenance, while large language models embedded in the pipeline adapt to changing website structures and assess data relevance. Results show that this architecture not only overcomes the limitations of traditional scrapers but also establishes a reusable framework for enterprise-scale knowledge extraction and real-time information access.

Building on this foundation, the thesis introduces a Retrieval-Augmented Generation (RAG) chatbot tailored for auto repair and appliance service businesses. Rather than serving solely as a validation tool, the RAG system constitutes a core contribution by transforming raw scraped data into a structured, queryable knowledge base that supports natural language interactions. The system leverages this structured knowledge to deliver accurate, domain-specific responses, demonstrating the practical utility of combining intelligent scraping with conversational AI.

Together, the multi-agent system and RAG chatbot highlight how intelligent agent architectures can advance both web data extraction and domain-specific conversational AI.

Keywords— Multi-agent systems, Web scraping, knowledge extraction, Retrieval-Augmented Generation (RAG), conversational AI, large language models, intelligent agents, data pipeline, semantic indexing

Résumé

Les méthodes traditionnelles de Web scraping rencontrent d'importants problèmes de scalabilité et de maintenance lorsqu'elles sont confrontées aux sites modernes riches en JavaScript, aux mécanismes anti-bots et aux contenus dynamiques. Cette thèse propose un système multi-agents qui automatise le Web scraping à grande échelle et l'extraction de données en répartissant les tâches entre des agents spécialisés responsables de l'exploration d'URL, du rendu de contenu, de l'analyse des données et de l'indexation sémantique.

Le cadre proposé assure la scalabilité grâce à l'exécution parallèle des tâches, à une gestion robuste des erreurs et à une maintenance modulaire, tandis que les modèles de langage de grande taille intégrés au pipeline s'adaptent aux évolutions des structures de sites Web et évaluent la pertinence des données. Les résultats montrent que cette architecture dépasse les limites des scrapers traditionnels et établit un cadre réutilisable pour l'extraction de données à l'échelle industrielle et l'accès à l'information en temps réel.

Sur cette base, la thèse introduit un chatbot à génération augmentée par la recherche (Retrieval-Augmented Generation, RAG) adapté aux entreprises de réparation automobile et d'électroménagers. Plutôt que de se limiter à un simple outil de validation, le système RAG constitue une contribution centrale en transformant les données brutes extraites en une base de données structurée et interrogeable, permettant des interactions en langage naturel. Ce système exploite ainsi ces données structurées pour fournir des réponses précises et spécifiques au domaine, démontrant l'utilité pratique de la combinaison entre scraping intelligent et intelligence conversationnelle.

Ensemble, le système multi-agents et le chatbot RAG mettent en évidence la manière dont les architectures à agents intelligents peuvent faire progresser à la fois l'extraction de données Web et l'IA conversationnelle spécialisée.

Mots clés systèmes multi-agents, Web scraping, extraction de connaissances, RAG, IA conversationnelle, grands modèles de langage, agents intelligents, pipeline de données, indexation sémantique.

الملخص

تواجه طرق استخراج البيانات التقليدية من الويب مشاكل كبيرة في التجميع والصيانة عند التعامل مع المواقع الحديثة الغنية بـ [جافا سكريبت]، والآليات المضادة للروبوتات، والمحتوى الديناميكي. تقترح هذه الأطروحة نظاماً متعدد الوكلاء يقوم بأتمتة استخراج البيانات من الويب على نطاق واسع واستخلاص المعرفة من خلال توزيع المهام بين وكلاء متخصصين مسؤولين عن استكشاف روابط الويب، عرض المحتوى، تحليل البيانات، والفهرسة الدلالية.

يضمن الإطار المقترح إمكانية التوسع عبر التنفيذ الموازي للمهام، وإدارة الأخطاء بشكل قوي، والصيانة المعيارية، بينما تتكيف نماذج اللغة الكبيرة المدججة في مسار البيانات مع تطورات هياكل المواقع وتقييم مدى ملاءمة البيانات. تُظهر النتائج أن هذه البنية لا تتجاوز فقط قيود أدوات الاستخراج التقليدية، بل تضع أيضاً إطاراً قابلاً لإعادة الاستخدام لاستخلاص المعرفة على نطاق المؤسسات والوصول للمعلومات في الوقت الحقيقي.

بناءً على هذا الأساس، تقدم الأطروحة روبوت محادثة يعتمد على التوليد المعزز بالاستخراج مخصص لأعمال إصلاح السيارات وخدمات الأجهزة الإلكترونية. بدلاً من أن يكون أداة تحقق فقط، يشكل النظام مساهمة جوهرية من خلال تحويل البيانات الخام المستخرجة إلى قاعدة معرفة منظمة وقابلة للاستعلام تدعم التفاعلات باستخدام اللغة الطبيعية. يستخدم النظام هذه المعرفة المنظمة لتوفير إجابات دقيقة ومتخصصة في المجال، مما يبرهن على الفائدة العملية للجمع بين الاستخراج الذكي والذكاء الاصطناعي الحوارية.

معاً، يبرز النظام متعدد الوكلاء وروبوت المحادثة كيف يمكن لبنيات الوكلاء الذكية تعزيز كل من استخراج بيانات الويب والذكاء الاصطناعي الحوارية المتخصص في المجال.

الكلمات المفتاحية: أنظمة متعددة الوكلاء، استخراج البيانات من الويب، استخلاص المعرفة، التوليد المعزز بالاستخراج، الذكاء الاصطناعي الحوارية، نماذج اللغة الكبيرة، الوكلاء الأذكياء، مسار البيانات، الفهرسة الدلالية.