OBIETTIVO DATI AMBIENTALI PIÙ ACCESSIBILI A TUTTI

L'INTELLIGENZA ARTIFICIALE CONVERSAZIONALE HA UN ENORME POTENZIALE PER RENDERE I DATI AMBIENTALI DI ISPRA E SNPA PIÙ FRUIBILI PONENDO DOMANDE IN LINGUAGGIO NATURALE. QUALI SONO GLI STRUMENTI DISPONIBILI E DA SVILUPPARE? QUALI LE SFIDE PER OTTENERE SOLUZIONI EFFICACI E RISPOSTE AFFIDABILI?

Attualmente, Ispra[1] e Snpa[2] offrono una vasta gamma di banche dati ambientali. Tuttavia, per un utente non esperto, consultare e correlare questi dati può risultare complesso e dispendioso in termini di tempo. Ad esempio, un cittadino che volesse conoscere, per una data località, le variazioni della temperatura dell'aria e collegarle alla qualità dell'aria, dovrebbe consultare differenti basi di dati tematiche e utilizzare strumenti diversi per interpretare e correlare i dati (*figura 1*).
A semplificare tutto ciò potrebbe essere di aiuto l'intelligenza artificiale conversazionale, grazie alla sua capacità di comprendere il linguaggio naturale e di estrarre e sintetizzare informazioni da enormi dataset, trasformando una ricerca frammentata in un'esperienza più fluida e accessibile. Lo scenario auspicabile è quello in cui un cittadino, uno studente o un ricercatore possa semplicemente porre domande in linguaggio naturale e ricevere risposte chiare e accurate, anche su temi scientifici complessi.

L'integrazione degli strumenti

Large language model (Llm) e *Retrieval-augmented generation* (Rag) sono alcuni strumenti di intelligenza artificiale che possono realizzare questa visione.
Il 2020 ha segnato una svolta significativa con il rilascio di Gpt-3 di OpenAi[3], un Llm che ha rivoluzionato il campo della comprensione e generazione del linguaggio naturale. Basati sull'architettura di *deep learning Transformer*[4], questi modelli, addestrati su vaste collezioni di testi, hanno dimostrato capacità inaspettate e sorprendenti in vari contesti, dall'analisi statistica alla creazione di rappresentazioni grafiche.
Tuttavia, nonostante il loro potenziale, gli Llm presentano alcune limitazioni, come la possibilità di generare informazioni inesatte o obsolete, soprattutto in settori specialistici o in rapida evoluzione, e possono confondere termini simili, utilizzati in contesti diversi, producendo risposte imprecise.
Per superare questi limiti, è stato sviluppato il *framework* Rag che combina le capacità di un Llm con l'accesso a una base di conoscenza specifica. Il funzionamento di Rag si basa su due componenti principali: un *retriever*, che individua e raccoglie le informazioni pertinenti da fonti specifiche, come database o raccolte documentali, e un generatore che utilizza quanto fornito dal *retriever* per produrre risposte in linguaggio naturale.

In sintesi, questa combinazione consente di avere risposte più accurate e pertinenti, garantendo che queste siano sempre basate su informazioni aggiornate e autorevoli anche in contesti complessi, senza la necessità di ricostruire l'intero modello linguistico.

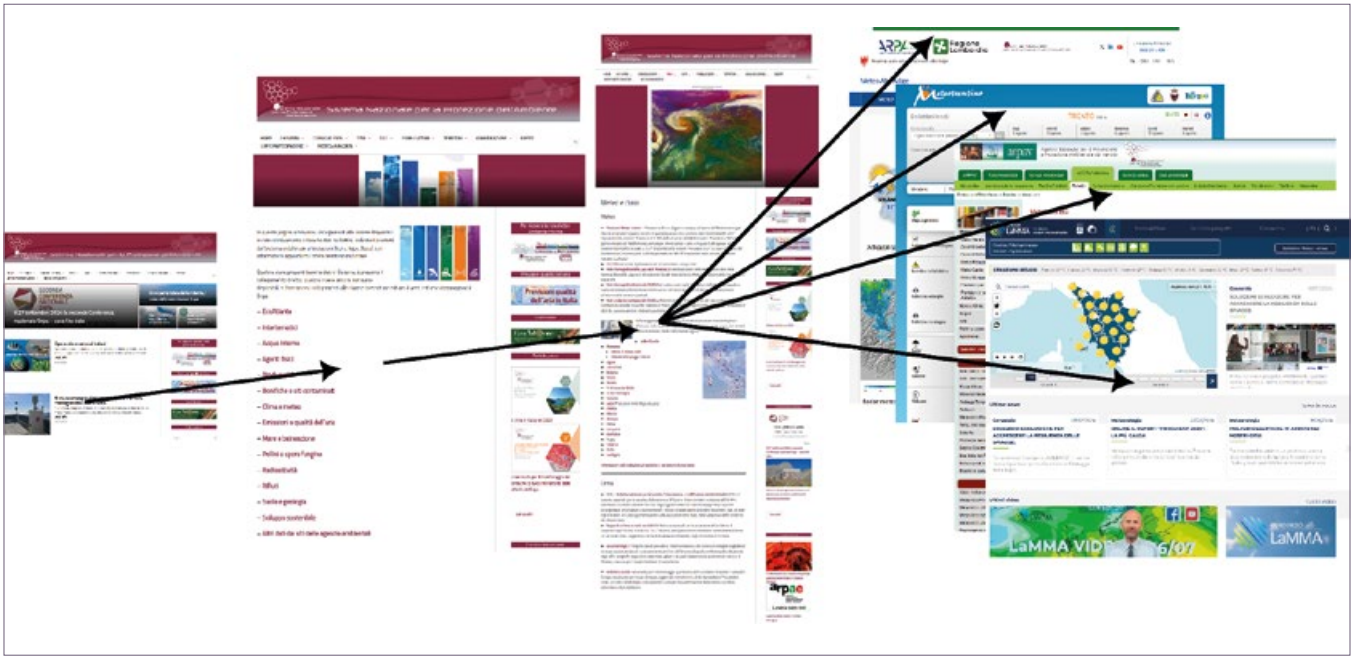

IMMAGINE: G. SCATENA - S. CIATTONI - CC BY-NC-SA 4.0

FIG. 1 RICERCA DATI CLIMATICI
Rappresentazione del percorso di un utente alla ricerca di dati climatici per un confronto tra dati di regioni diverse.

Metodologie implementative e sfide per i dati ambientali

Per realizzare lo scenario auspicato in precedenza è necessario quindi prevedere le seguenti componenti: un sistema Llm, un Rag con relativa base dati vettoriale e un'interfaccia utente, tipicamente un chatbot.
Nel seguito, descriviamo alcune metodologie implementative disponibili a oggi, evidenziando pro e contro nell'ambito di applicazione al dominio dei dati ambientali.
Gli strumenti recentemente sviluppati e attualmente disponibili in questo ambito variano in base alle seguente principali caratteristiche:
- livello di maturità e semplicità di utilizzo: vi sono approcci che richiedono lo sviluppo e la scrittura di codice[5], mentre stanno emergendo *framework* che consentono di creare il proprio chatbot tramite installazione semplificata e interfaccia grafica di creazione della base di conoscenza
- libertà di scelta del generatore Llm: vi sono approcci che utilizzano esclusivamente Llm di terze parti, tipicamente *cloud* a pagamento, mentre altri consentono di personalizzare il motore utilizzato, aprendo la possibilità di utilizzare modelli *open source* con esecuzione su risorse private. L'approccio *open source*[6] ha vantaggi in termini di *privacy* dei dati, trasparenza, controllo e indipendenza, personalizzazione e integrazione, ma è necessario valutarne attentamente la scalabilità e i costi a lungo termine delle risorse *hardware*
- specializzazione del *retriever* in un ambito applicativo: ne esistono infatti di specializzati nell'analisi di documenti, con utilizzo di riconoscimento ottico dei caratteri su scansioni, o su dati tabellari.

Abbiamo testato alcuni *framework open source* (*tabella 1*) con campioni di dati ambientali provenienti dalle basi di dati Ispra e Snpa.
I risultati ottenuti sono promettenti ma evidenziano la necessità di *hardware* dedicato per l'esecuzione di modelli Llm su infrastrutture private.
Nonostante ciò, i prototipi sviluppati dimostrano già la capacità di fornire risposte pertinenti e di citare fonti rilevanti (*figure 2 e 3*).
Tuttavia, le tecnologie Llm e Rag applicate ai dati ambientali presentano ancora alcune limitazioni dovute alle caratteristiche dei dati stessi e spesso vengono fornite risposte incomplete o incoerenti, a confronto con i risultati

Approccio	Llm	Codice sorgente	Demo
RAGFlow	Molteplici	https://github.com/infiniflow/ragflow	https://demo.ragflow.io/
Cognita	Molteplici	https://github.com/truefoundry/cognita	https://cognita.truefoundry.com/
Quivr	Molteplici	https://github.com/QuivrHQ/quivr	https://chat.quivr.app/
Oraculum	ChatGpt	https://github.com/Unipisa/Oraculum	https://sibyllae.unipi.it/missioni

TAB. 1 FRAMEWORK
Strumenti open-source testati con campioni di dati ambientali.

Fonte: Banche dati Ispra e Snpa

IMMAGINE: G. SCATENA - S. CIATTONI - CC BY-NC-SA 4.0

FIG. 2 QUIVR
Un esempio di creazione di un sistema con Quivr: una volta aggiunta la conoscenza (a) è possibile interrogarla in linguaggio naturale in modo conversazionale (b).

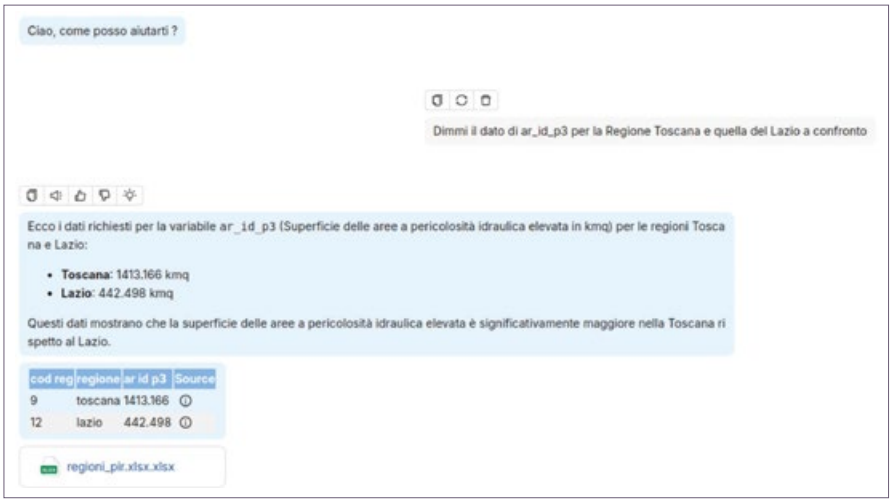

IMMAGINE: G. SCATENA - S. CIATTONI - CC BY-NC-SA 4.0

FIG. 3 RAGFLOW
Esempio di interazione con RagFlow: l'agente, a cui è stata creata una base di conoscenza costituita unicamente da fogli di calcolo, comprende la domanda, seleziona e cita la fonte dati pertinente, restituisce il valore desiderato estraendola da essa, spiegandone il significato ed elaborando una valutazione contestuale.

ottenuti dalla loro applicazione in altri domini applicativi[7].
Le caratteristiche dei dati ambientali che rendono più complessa l'applicazione di Llm e Rag rispetto ad altri domini sono molteplici. I dati sono spesso frammentati, provenienti da diverse fonti con formati e schemi non standardizzati, complicando l'integrazione e l'analisi; le relazioni causali nei dati ambientali sono complesse e contestuali, rendendo difficile, per i modelli linguistici, cogliere correttamente le connessioni tra fenomeni; il linguaggio tecnico e specialistico utilizzato è spesso difficile da interpretare per modelli addestrati su testi generici. Infine, la natura non strutturata e disomogenea di questi dati rappresenta un'ulteriore sfida per la loro comprensione ed elaborazione.
Per ottenere una soluzione efficace al problema è quindi necessario:
- sviluppare *retriever* adatti al dominio applicativo, unendo l'utilizzo di approcci basati su *embedding* (capacità di comprensione di dati testuali non strutturati) e approcci basati su grafi (capacità di rappresentare relazioni complesse tra entità, adatti per analizzare dati spaziali e temporali)
- unificare e centralizzare dati e documenti: la qualità dei dati, ovvero un dataset ben organizzato e strutturato, assieme a un corpus documentale privo di lacune, facilita il lavoro dei *retriever*, migliorando la precisione e la velocità delle risposte.

Conclusioni

Le tecnologie emergenti che combinano Llm e Rag offrono un promettente futuro per l'interazione uomo-macchina e potrebbero rivoluzionare la fruibilità dei dati soprattutto in ambiti come l'analisi di dati ambientali.
L'utilizzo di Rag riesce a risolvere alcuni problemi propri degli Llm come l'affidabilità delle risposte in determinati ambiti, conservando le modalità di interazione discorsiva e conscia del contesto tipiche dei Llm.

Nonostante i notevoli progressi, persistono sfide legate alla gestione di dati eterogenei e non strutturati, nonché all'elevato costo computazionale. Per un'applicazione efficace ai dati ambientali, è necessaria una standardizzazione dei dati e un approfondimento delle tecniche di addestramento dei modelli Rag su dati semi-strutturati di tipo non testuale. Crediamo fermamente nel potenziale di queste tecnologie per rendere i dati ambientali più accessibili e auspichiamo un maggiore impegno da parte della comunità scientifica e industriale.

Guido Scatena, Simona Ciattoni

Istituto superiore per la protezione e la ricerca ambientale (Ispra)

NOTE

[1] www.isprambiente.gov.it/it/banche-dati

[2] www.snpambiente.it/dati/

[3] https://openai.com/

[4] Vaswani A., Shazeer N., Parmar N., Uszkoreit J., Jones L., Gomez A.N., Kaiser L., Polosukhin I., 2017, "Attention is all you need", Nips'17: Proceedings of the 31st International Conference on Neural Information Processing Systems, https://arxiv.org/abs/1706.03762

[5] Per poter sviluppare strumenti *ad hoc* è necessario conoscere librerie come PyTorch (https://pytorch.org), TensorFlow (www.tensorflow.org) e le librerie di generazione (Trasformers, https://huggingface.co/docs/transformers/llm_tutorial) in continuo sviluppo su HuggingFace (https://huggingface.co).

[6] Si vedano, ad esempio, quelli disponibili su https://ollama.com

[7] Ad esempio, l'utilizzo di Oraculum con Rag sul corpus documentale di amministrazione Università di Pisa, utilizzabile tramite Sibylla (https://sibyllae.unipi.it/missioni), fornisce risposte molto soddisfacenti.

REGOLAMENTO UE 2024/1689

LA LEGISLAZIONE EUROPEA A TUTELA DEI RISCHI DELL'INTELLIGENZA ARTIFICIALE

In Europa la legge sull'intelligenza artificiale (Ai Act) è entrata in vigore il 1° agosto 2024. Alcune disposizioni sono già pienamente applicabili mentre altre necessitano di un periodo transitorio a causa della loro maggiore complessità e di requisiti da adottare.

La Commissione europea sta promuovendo il *Patto sull'intelligenza artificiale,* un'iniziativa volontaria per supportarne l'implementazione e invitare gli svilppatori ad adottare quanto previsto nell'Ai Act prima delle scadenze. Il patto si struttura su due pilastri:

Primo pilastro: raccolta e scambio con la rete del Patto sull'intelligenza artificiale
Ha il ruolo di essere il punto di accesso per coinvolgere la rete del patto per l'intelligenza artificiale costituita da organizzazioni interessate al patto; incoraggia lo scambio di migliori pratiche tramite la condivisione di esperienze e conoscenze. Fornisce inoltre informazioni sul processo di applicazione della legge dell'intelligenza artificiale.

Secondo pilastro: facilitare e comunicare gli impegni aziendali
Il suo scopo è fornire un quadro per promuovere la rapida attuazione di alcune delle misure della legge sull'intelligenza artificiale. Incoraggia i fornitori e gli operatori di sistemi di intelligenza artificiale a prepararsi tempestivamente e ad adottare misure per conformarsi ai requisiti e agli obblighi stabiliti dalla legge.

Il quadro normativo definisce quattro livelli di rischio per i sistemi di intelligenza artificiale:

1) rischio minimo
2) rischio limitato, sistemi di intelligenza artificiale con specifici obblighi di trasparenza
3) alto rischio
4) rischio inaccettabile.

La legge europea sull'intelligenza artificiale è il primo quadro giuridico a livello mondiale che affronta i rischi dell'la garantendo che i sistemi di la rispettino i diritti fondamentali, la sicurezza e i principi etici affrontando i rischi di modelli di la molto potenti e di grande impatto. (*DM*).

Fonte: Commissione europea (https://digital-strategy.ec.europa.eu/it/policies/regulatory-framework-ai)
Regolamento Ue 2024/1689 https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32024R1689