

PROGETTI CONCLUSI 2019 II Call

Titolo progetto	Il web scraping e la domanda di lavoro
Descrizione	<p>In ambito europeo, è in corso un progetto implementativo sulle Online Job Vacancies in cui l'Italia è coinvolta, all'interno del più ampio progetto ESSnet on Big Data II.¹</p> <p>I Paesi Membri partecipanti hanno concordato che gli annunci di lavoro online non possano essere utilizzati per sostituire le statistiche ufficiali esistenti sui posti vacanti (come richieste dai Regolamenti europei). Tuttavia, la granularità dell'informazione in essi contenuta può fornire utili approfondimenti non derivabili dalle statistiche ufficiali correnti.</p> <p>E' reputato, dunque, importante valutare la possibilità di utilizzare i dati di tali annunci per arricchire, mediante indicatori supplementari, le statistiche ufficiali attualmente diffuse sui posti di lavoro vacanti.</p> <p>In tale contesto, nell'ambito di questo progetto, è stato fatto, in prima battuta web scraping degli annunci di lavoro online e poi dal risultato del web scraping, è stato estratto il testo utile alla classificazione degli annunci, secondo la tipologia della professione (secondo la classificazione ufficiale delle professioni ISTAT CP2011) e la localizzazione della posizione lavorativa pubblicizzata.</p>
Obiettivi	Il progetto ha l'obiettivo di sviluppare metodi, procedure e raccomandazioni di supporto alla raccolta, trattamento e analisi degli annunci di lavoro online ai fini della produzione di statistiche supplementari di supporto alle statistiche ufficiali sui posti di lavoro vacanti, attualmente diffuse.
Metodologia	Le attività hanno riguardato la definizione dei metodi di web scraping del testo degli annunci, quindi la scelta dei siti web fra tutti quelli che pubblicano annunci di lavoro, e delle tecniche di estrazione e archiviazione delle informazioni con cui nei vari siti vengono descritti gli annunci di lavoro. Tali dati sono poi stati sottoposti ad operazioni di data cleaning e di standardizzazione. E' stata, inoltre, studiata e implementata una metodologia per identificare e classificare, all'interno del testo dell'annuncio, le informazioni descrittive delle caratteristiche della posizione lavorativa pubblicizzata. In particolare, attraverso delle tecniche di Machine Learning, ogni annuncio è stato classificato sulla base delle informazioni contenute nel suo testo, in base sia alla localizzazione sia alla professione della posizione lavorativa (secondo la classificazione ufficiale delle professioni ISTAT CP2011).
Risultati ottenuti	Nel report finale del progetto, sono stati descritti tutti i limiti di rappresentatività dei dati sugli annunci pubblicati sui portali. Infatti, tali annunci di ricerca di candidati all'assunzione possono al più coprire i posti vacanti per i quali la ricerca attiva avvenga almeno anche tramite questo canale, che rappresenta un sottoinsieme non necessariamente rappresentativo del totale dei posti vacanti così come definiti dal regolamento di riferimento. Inoltre, l'attività di web
specificare l'impatto sulla produzione statistica	

¹ Successivo ad ESSnet Big Data, svoltosi da gennaio 2016 a maggio 2018. Tale progetto è, più in generale, finalizzato all'integrazione dei big data come fonti nella produzione della statistica ufficiale.

scraping svolta nell'ambito del progetto ha riguardato solo i più importanti portali di offerta di lavoro. Di conseguenza, gli annunci ricavati da tale attività potrebbero non coprire nemmeno tutti i posti vacanti per cui si utilizza almeno anche questo canale.

Tuttavia, le analisi successive hanno mostrato che caratteristiche importanti delle posizioni lavorative pubblicizzate possono essere estratte dalle informazioni contenute negli annunci di lavoro online, in particolare: tipo di professione, regione, tipo di contratto e orario di lavoro.

Nonostante i limiti indicati, questi dati rappresentano, dunque, una rilevante fonte di potenziale arricchimento dell'informazione statistica ufficiale attualmente diffusa sulla domanda di lavoro.

Al momento, infatti, le statistiche ufficiali sui posti vacanti diffuse in Italia e trasmesse a Eurostat² non sono disaggregate per professione o regione. Tuttavia, entrambe queste disaggregazioni sono considerate come molto utili dagli utenti attuali e potenziali di tali statistiche.

² Tali statistiche sono ottenute tramite l'integrazione dei micro-dati dell'Indagine trimestrale sui posti vacanti e le ore lavorate (Vela), con quelli della Rilevazione mensile sull'occupazione, gli orari di lavoro, le retribuzioni e il costo del lavoro nelle grandi imprese (GI), e l'utilizzo come fonte ausiliaria dei micro- e macro-dati della Rilevazione trimestrale su Occupazione Retribuzioni e Oneri Sociali (Oros).