istat
working
papers

N.11
2019

Ricerca e codifica nelle classificazioni della statistica ufficiale

Alessandro Capezzuoli

Istat
Istituto Nazionale
di Statistica

Istat Working Papers

Ricerca e codifica nelle classificazioni della statistica ufficiale

N. 11/2019

ISBN 978-88-458-1997-1

Ricerca e codifica nelle classificazioni della statistica ufficiale

Alessandro Capezzuoli[1]

Sommario

Il presente lavoro ha l'obiettivo di illustrare un insieme di metodi e tecniche per eseguire la ricerca e la codifica all'interno delle classificazioni statistiche. Le soluzioni teoriche volte a rispondere alle criticità connesse ai sistemi di ricerca e le relative soluzioni tecnologiche adottate forniscono una metodologia completa – generalizzabile ad ogni classificazione – per la costruzione di strumenti efficaci che possano essere condivisi e disseminati all'interno di applicazioni web e questionari statistici, attraverso web service RESTful, widget e applicazioni ad hoc. Elastisearch è lo strumento utilizzato per effettuare la ricerca semantica all'interno delle classificazioni e dei dati relativi all'indagine sulle professioni. Questo tipo di ricerca non esaurisce tutti i possibili metodi di codifica: all'interno del documento verranno analizzate anche le possibilità offerte dai questionari interattivi e dai test di autovalutazione.

Parole chiave: classificazioni ufficiali, motori di ricerca semantici, diffusione dati, *data sharing*, codifica.

Abstract

The aim of the present work is to illustrate a set of methods and techniques to perform search and coding within statistical classifications. In order to meet the critical issues related to the search systems, theoretical and technological solutions providing a complete methodology – generalizable to each classification – for the construction of effective tools for coding are proposed. These tools can be applied to most of the official classifications and be shared and disseminated within web applications and statistical questionnaires through RESTful web services, widgets and ad hoc applications. Elasticsearch is a distributed, RESTful search and analytics engine capable of perform and combine many types of searches within taxonomies and textual documents. Semantic search is just one of many coding possibilities: the opportunities offered by interactive questionnaires and self-assessment tests are also analyzed in this document.

Keywords: official classifications, semantic search engines, data dissemination, data sharing, coding.

[1] Le opinioni espresse riguardano esclusivamente l'autore e non implicano alcuna responsabilità da parte dell'Istat.
Si ringrazia la dottoressa Emanuela Recchini per i numerosi contributi nella stesura e nella revisione.

Indice

Pag.

Introduzione

Le classificazioni statistiche sono fondamentali per la produzione di statistiche ufficiali, per il raccordo di dati provenienti da diverse fonti e per la creazione di un linguaggio standardizzato che consenta la comparabilità dei dati, il *linkage* e la descrizione dei fenomeni. La problematica principale legata all'uso delle classificazioni è la codifica, ovvero la corretta attribuzione della coppia chiave-descrizione rispetto a una ricerca eseguita sulla base di un criterio specifico (es. albero decisionale, ricerca testuale, etc.). Le classificazioni statistiche possono essere utilizzate in molteplici ambiti, ad esempio nella ricerca, all'interno di indagini campionarie e censuarie, all'interno di archivi amministrativi o di sistemi informativi distribuiti che collegano tra loro dati provenienti da diverse fonti. Gli utilizzatori delle classificazioni possono avere esigenze molto diverse; basti pensare a un intervistatore che deve individuare una certa professione nel corso di un'indagine, a un utente che deve indicare il codice dell'attività economica di un'impresa o a un utente che cerca delle informazioni collegate ad un determinato titolo di studio. La probabilità di una codifica errata può essere molto alta e può variare rispetto alla conoscenza che l'utilizzatore ha della logica e del linguaggio adottato dalla classificazione.

In questo documento saranno illustrate le principali tecniche di codifica e le soluzioni tecnologiche utilizzate per semplificare la ricerca all'interno dei sistemi classificatori e supportare i diversi utenti (intervistatori, rispondenti, istituzioni, etc.). Si focalizzerà l'attenzione sui punti di forza e di debolezza della codifica manuale e sulle possibili applicazioni degli strumenti offerti dal web e connessi a questa tecnica, che prevede l'interazione diretta di un sistema informativo con l'utente. Non saranno presi in esame i processi di codifica automatica, che implicano l'attribuzione probabilistica di un elemento rispetto ad una stringa testuale (è il caso, ad esempio, della codifica delle cause di morte, che viene eseguita sulla base di un algoritmo che interpreta la descrizione fornita dal medico che ha accertato il decesso e associa ad essa il codice più appropriato).

Nei capitoli che seguono, dopo una panoramica sulle caratteristiche fondamentali delle tassonomie e le criticità che la codifica manuale comporta, si entrerà nel merito dei principali strumenti, delle tecniche e delle tecnologie finalizzate alla codifica. Nel prosieguo, il termine tassonomia verrà usato per indicare le classificazioni estensionali, ovvero quei tipi di classificazioni in cui gli oggetti di un insieme vengono raggruppati in due o più sottoinsiemi in modo da massimizzare la somiglianza fra membri dello stesso sottoinsieme e la dissomiglianza fra membri di sottoinsiemi diversi. La maggior parte delle classificazioni statistiche ufficiali fanno parte delle tassonomie o classificazioni estensionali. Un focus particolare riguarderà alcuni casi di analisi specifici – connessi alla Classificazione delle Professioni (CP2011) e alla Classificazione Statistica Internazionale delle Malattie e dei Problemi Sanitari Correlati (International Classification of Diseases, ICD-10) – all'interno dei quali verrà proposto un insieme di soluzioni ad hoc per integrare la ricerca testuale. Infine, saranno illustrati i principali strumenti di condivisione e disseminazione relativi alle classificazioni statistiche.

2 Le classificazioni statistiche

2.1 Tassonomie e classificazioni

Una tassonomia (dal greco: τάξις, taxis, ordinamento e νόμος, nomos, norma o regola) è un insieme ordinato di oggetti messi in correlazione tra loro. Le tassonomie sono delle strutture concettuali che "ritagliano" un certo ambito di conoscenza attraverso l'intensione e l'estensione del concetto che rappresentano. *Il concetto è un ritaglio operato in un flusso di esperienze infinito in estensione e in profondità ed infinitamente mutevole. Il ritaglio si opera considerando globalmente un certo ambito di queste esperienze: ad esempio, unificando alcune sensazioni visive e tattili nel concetto di "tavolo" oppure alcuni stati d'animo nel concetto di rabbia. Effettuato una volta questo conglobamento di sensazioni, ci sarà più facile ripeterlo in casi analoghi, per cui riconosceremo (non senza margini di errore) altri tavoli o altri stati di rabbia. In questa maniera ridurremo gradatamente la complessità e la problematicità del mondo esterno, e quindi accresceremo la no-*

stra capacità di orientamento nella realtà." (Marradi, 1992). Semplificando al massimo, si può dire che l'intensione di un concetto è l'insieme delle caratteristiche che lo differenziano dagli altri. L'estensione, invece, rappresenta il numero degli oggetti a cui si può applicare un concetto. L'intensione e l'estensione sono inversamente proporzionali: all'aumentare della prima diminuisce la seconda.
Per chiarire questa relazione, si prenda ad esempio il concetto di "pianta":

Intensione: acquatica, verde, con fiori rampicante;
Estensione: tutte le piante sono oggetti che fanno parte dell'intensione;

Se si aumenta l'intensione, aggiungendo ad esempio "pianta medicinale", è evidente la conseguente riduzione dell'estensione.
Le tassonomie, o classificazioni estensionali, hanno degli elementi comuni che ne definiscono la logica e la struttura:

1) sono costruite a partire da più aspetti dell'intensione del concetto di genere (*fundamentum divisionis*); a seconda di quanti aspetti articolano, si ottengono strutture concettuali più semplici (classificazione), complesse (tipologie) e molto complesse (tassonomie);
2) le classi, i tipi e le tassonomie devono rispettare il criterio della mutua esclusività;
3) le classi, i tipi e le tassonomie devono rispettare il criterio della esaustività.

Le tassonomie sono essenzialmente delle classificazioni all'interno delle quali le divisioni dell'estensione sono operate in successione su concetti di generalità decrescente. Il *fundamentum divisionis* è l'aspetto dell'intensione del concetto generale, che viene articolato per formare i vari concetti di classe, e distingue la tassonomia dalle altre forme di divisione di un insieme. È utile sottolineare che alcune classificazioni, perlopiù relative ai fenomeni sociali, presentano dei vulnus a causa di una errata definizione del *fundamentum*.
La mutua esclusività delle tassonomie prevede che ciascun oggetto non possa essere attribuito a più istanze (gruppi di concetti). L'esaustività implica che ogni oggetto debba essere attribuito almeno a una istanza.

I concetti, pur essendo ideologicamente indipendenti dai termini, nella realtà sono strettamente legati ad essi. La conoscenza dell'insieme dei termini presenti all'interno di una tassonomia è fondamentale per la definizione di metodi e soluzioni finalizzate alla ricerca semantica.

2.2 Codifica manuale e sistemi di ricerca

La codifica manuale prevede che un utente interagisca direttamente con la tassonomia, attraverso un sistema di consultazione, che solitamente restituisce un insieme di risultati in funzione di un input specifico (ricerca testuale, ricerca di un codice, consultazione di un albero gerarchico, etc.). In questo caso, l'approccio di tipo probabilistico potrebbe essere fuorviante e fortemente condizionante.

I sistemi di codifica, ormai generalmente presenti in ambiente web, molto spesso consentono di eseguire ricerche differenziate attraverso diverse tecniche allo scopo di ridurre l'errore, ottimizzare i tempi e migliorare la qualità del risultato. Una codifica di qualità è associata essenzialmente all'individuazione, da parte degli utenti, della classe più adeguata da utilizzare. Spesso, a causa dell'errata definizione del *fundamentum divisionis* e di un sistema di codifica inadeguato, gli utilizzatori scelgono delle classi appropriate. Le criticità legate a qualsiasi sistema di ricerca sono sostanzialmente tre: i falsi risultati positivi, i falsi risultati negativi e la diversità degli utenti. Per i primi due aspetti è possibile adottare delle soluzioni logiche e tecnologiche che migliorino considerevolmente il ventaglio di risultati proposti: ridurre il numero dei falsi risultati positivi e proporre dei sistemi di ricerca che consentano di individuare i falsi risultati negativi permette di migliorare notevolmente la codifica. La diversità tra gli utenti, invece, è un problema di difficile soluzione poiché riguarda i per-

corsi logici che vengono eseguiti da ciascun soggetto durante una ricerca, il livello di conoscenza della lingua, le diverse tecniche di memorizzazione e gli strumenti disponibili.

Negli ultimi anni, il web e i motori di ricerca hanno condizionato fortemente il modo di cercare le informazioni ed hanno "educato" gli utenti ad usare le parole nel modo giusto e a comporle efficacemente per ottenere dei risultati soddisfacenti. Un motore di ricerca, semplificando al massimo l'enorme complessità degli algoritmi, utilizza il potere discriminante che hanno le parole all'interno di un insieme di documenti/record. L'output è una lista di risultati ordinati sulla base di un certo punteggio. Una parola può avere un potere discriminante diverso in base ai documenti da analizzare: in un archivio che contiene soltanto documenti scientifici e un solo documento legato alla narrativa, la parola "narrativa" avrà un altissimo potere discriminate rispetto alla situazione inversa. Google è un enorme contenitore di documenti di ogni tipo e il suo sistema di ricerca prevede numerosi accorgimenti per migliorare la precisione e l'efficacia: basti pensare alla diversità dei risultati che si ottengono effettuando una ricerca di una frase delimitata o meno dai doppi apici. Nel primo caso, il potere discriminante della sequenza esatta di parole è sicuramente più alto di quello di ogni singola parola.

C'è da dire che, anche nel caso di una ricerca senza i doppi apici, molti algoritmi prevedono una suddivisione delle frasi in singoli *token* lessicali, ovvero blocchi di testo costituiti da caratteri indivisibili, e un'analisi avanzata e pesata della loro distribuzione all'interno dei documenti da cercare. Il potere discriminante di una parola è inversamente proporzionale alla quantità di documenti in cui essa compare e alla relativa frequenza. Per questo è nato un filone di ricerca che si occupa della semantica, ovvero delle tecniche per individuare il collegamento logico tra una parola e un'altra, al fine di definire il senso di una frase (ontologia).

È utile sottolineare che, riferendosi a una classificazione statistica, la ricerca testuale non è l'unica possibile. Trattandosi perlopiù di strutture gerarchiche ad albero (costituite da nodi e archi/padri e figli), esistono altri metodi per individuare una foglia, partendo dalla radice. La ricerca testuale risponde a un gran numero di esigenze, ma non copre tutta la casistica possibile. In particolare, essa risulta poco efficace nei casi in cui occorre classificare un oggetto la cui descrizione non è presente nel dizionario della classificazione. Poiché una tassonomia deve rispettare il criterio dell'esaustività, i casi in cui non c'è una corrispondenza tra la parola cercata e i termini della tassonomia rientrano nella casistica dei falsi risultati negativi. Questa situazione può verificarsi molto frequentemente, a causa dell'ampio arco temporale di validità di una classificazione ufficiale (10/20 anni).

3 Strumenti, tecniche e tecnologie per la codifica

3.1 La ricerca testuale

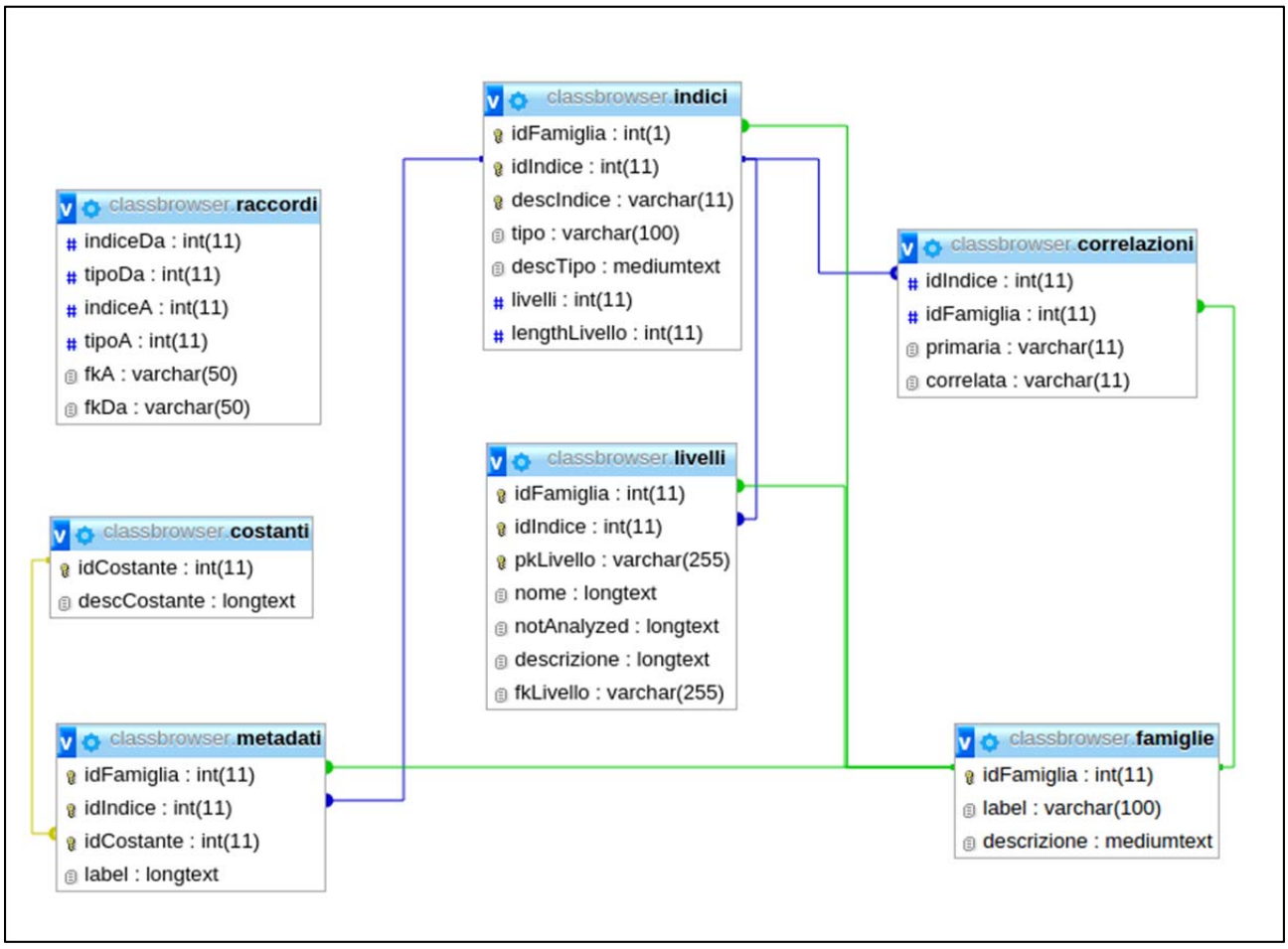
La ricerca testuale è una tecnica molto familiare agli utenti che cercano informazioni sul web. Per questo, non necessitando di particolari abilità che non siano state già acquisite attraverso la navigazione in internet, si presta anche per effettuare ricerche all'interno delle classificazioni statistiche. L'insieme dei termini usati nelle tassonomie per descrivere i concetti possono differire molto tra loro e possono discostarsi notevolmente dal linguaggio parlato. Ad esempio, la parola "impiegato" viene usata comunemente per identificare un ampio numero di professioni, ma all'interno della Classificazione Ufficiale delle Professioni corrisponde a poche e circoscritte unità professionali. La ricerca testuale di questa parola produce dei falsi risultati negativi, ovvero l'omissione di quegli elementi presenti all'interno della tassonomia – facenti parte dell'intensione "professioni impiegatizie" – che non vengono restituiti come risultati della ricerca in quanto la stringa cercata non è presente nel dizionario. Un problema analogo, sempre nell'ambito della Classificazione delle Professioni, è connesso alla parola "macellaio", che nel linguaggio comune viene associato all'esercente per le vendite nei negozi di macelleria, mentre all'interno della tassonomia identifica il personale che lavora all'interno dei macelli: in questo caso, la ricerca produrrà contemporaneamente un falso risultato positivo e un falso risultato negativo. La ricerca della parola "meccanico", invece, produce

una quantità considerevole di risultati che rendono difficile la codifica: è il caso dei falsi risultati positivi.

Questa casistica, che prende in esame il caso specifico delle professioni, è riscontrabile all'interno della maggior parte delle classificazioni statistiche ufficiali. Le tassonomie sono caratterizzate da una struttura gerarchica ad albero costituita da un certo numero di istanze riferite a un singolo nodo radice che definisce le proprietà delle istanze sottostanti. All'ultimo nodo, frequentemente, è associato un ulteriore livello di esempi che contiene un consistente numero di parole vicine al linguaggio parlato e utile nel supporto alla ricerca e alla codifica. Questa caratteristica comune consente di generalizzare i sistemi di ricerca, per estenderli facilmente a numerose casistiche.

La maggior parte delle classificazioni statistiche viene archiviata in una base dati relazionale, all'interno della quale si possono eseguire delle interrogazioni in linguaggio SQL che essenzialmente consentono di effettuare due tipi di ricerca: "exact match" e "full text". Di solito, alle query SQL sono associati degli algoritmi che eseguono dei controlli di sintassi delle stringhe, per eliminare le parole inutili alla ricerca (*stopwords*) o per trovarne la radice (*stemming*). Un esempio di schema Entità-Relazione della base dati generalizzata che consente l'archiviazione delle classificazioni ufficiali è riportato nella Figura 1.

Figura - Schema ER base dati relazionale



Fonte: Elaborazioni sulla base dello schema ER utilizzato per la creazione del database

Le relazioni e le proprietà delle entità di questo particolare schema si possono facilmente ricondurre al modello GSIM per le classificazioni statistiche. In particolare, la base dati è strutturata in:

- Famiglie – descrizioni e identificativi dell'ambito di conoscenza rappresentato dalla tassonomia (professioni, attività economiche, etc.).
- Livelli – gerarchie complete delle diverse classificazioni relative a una certa famiglia e a un determinato indice.
- Indici – informazioni di dettaglio della classificazione (numero di livelli gerarchici, etichetta, etc.).
- Correlazioni – gestione delle relazioni esistenti tra oggetti della stessa classificazione.
- Raccordi – gestione dell'allineamento con altre classificazioni o differenti edizioni della stessa classificazione.

A partire da questo modello, è possibile effettuare una ricerca testuale con gli strumenti forniti dal linguaggio SQL. Ad esempio, la ricerca full-text all'interno di un DBMS MySQL può essere sintetizzata dalla query:

*SELECT DISTINCT pkLivello, nome, descrizione, fkLivello from livelli WHERE MATCH (`nome`) AGAINST ('**STRINGA_CERCATA**' IN BOOLEAN MODE)*

Una ricerca simile presenta numerose criticità legate al *parsing* delle stringhe, ovvero al trattamento e all'analisi sintattica del testo, e ai risultati restituiti dalla query.

Prendendo ancora una volta in esame la CP2011, è evidente che, laddove ci sia una corrispondenza esatta tra la parola cercata (es. Elettrauto) e il record da individuare (es. 6.2.4.1.5 – Elettrauto), un'interrogazione simile è più che sufficiente per ottenere un risultato soddisfacente. Poiché, al contrario, la casistica e la varietà delle ricerche sono molto ampie, occorre prevedere un sistema che tenga conto di:

- Singolari e plurali
- *Stopwords*, ovvero parole ininfluenti per la ricerca che dipendono dal dizionario e dalla lingua
- Termini al maschile e al femminile
- Differenze tra una ricerca effettuata con una parola singola e attraverso una frase
- Maiuscole e minuscole
- Errori di ortografia

3.2 Elasticsearch: un sistema generalizzato per la ricerca testuale

Negli ultimi anni, sono nati numerosi strumenti open source che sfruttano le potenzialità della libreria Lucene per la ricerca full-text. Lucene è un progetto open source della Apache Software Foundation altamente scalabile e personalizzabile che effettua una indicizzazione strutturata dei documenti e permette ricerche testuali molto avanzate attraverso numerosi strumenti di *preprocessing* per il supporto di diverse lingue. Lucene permette di eseguire query differenziate (phrase queries, wildcard queries, proximity queries, range queries, etc.) all'interno dei campi presenti nell'indice. Un indice contiene una sequenza di documenti. Un documento è composto da una sequenza di campi, ovvero una sequenza di termini. Elasticsearch (www.elastic.co) è un potente motore di ricerca che può essere impiegato in un ambiente distribuito con più nodi e fornisce numerosi servizi RESTful utilizzabili attraverso flussi e query JSON.

L'indicizzazione può essere applicata alle classificazioni per effettuare ricerche semantiche complesse e differenziate. Le possibilità offerte da un'architettura che utilizza Elasticsearch sono molteplici. Ad esempio, è possibile sfruttare interamente le potenzialità del protocollo HTTP, e i metodi PUT, GET, POST e DELETE, attraverso un linguaggio DSL (Domain Specific Language) nel formato JSON. Questa soluzione permette di semplificare notevolmente la formulazione di query molto complesse e rende fruibile il sistema di ricerca a partire da qualsiasi linguaggio di programmazione.

L'architettura REST (Representational State Transfer), oltre a facilitare l'interrogazione del motore di ricerca, permette di gestire agevolmente gli indici e gli item attraverso le operazioni

CRUD (CREATE, READ, UPDATE, DELETE). L'indicizzazione attraverso Elasticsearch permette di manipolare grandi moli di dati, grazie alla gestione interna dei documenti completamente svincolata dalle basi dati relazionali e alla possibilità di creare cluster distribuiti. La creazione di un indice, ad esempio per la famiglia di classificazioni legate alle professioni, utilizzando PHP e l'estensione cUrl, è molto semplice:

```
$ curl -XPUT 'http://localhost:9200/professioni/' -d '{
    "settings" : {
        "number_of_shards" : 3,
        "number_of_replicas" : 2
    }
}'
```

Ad ogni indice è stata associata una mappatura del documento, che contiene l'edizione della classificazione (attraverso la valorizzazione dell'opzione *type*), i campi, la loro tipologia e il tipo di *analyzer* sintetizzato dall'array JSON seguente:

```
PUT professioni
{
  ""mappings": {
    "CP2011": {
      "properties": {
        "nome": {  "type": "string",  "analyzer": "standard" }
      }
    }
```

L'*analyzer* è composto dal *tokenizer*, ovvero lo strumento che suddivide il testo in singoli *token*, e da un insieme di filtri attraverso i quali stabilire i criteri per effettuare il *parsing* della stringa di ricerca. Il popolamento dell'indice si effettua attraverso un *river*, ovvero un *plugin* che consente la connessione diretta, e la conseguente indicizzazione automatica, tra il database MySQL ed Elasticsearch. La creazione, la configurazione e il popolamento dell'indice consentono di ottenere un set di API REST da interrogare attraverso delle query string. Le API (Application Programming Interface) sono delle applicazioni software che consentono la diffusione di dati e metadati sul web. In particolare, le API REST (Representational State Transfer) permettono di sfruttare le potenzialità del protocollo HTTP, favorendo l'integrazione e lo scambio delle informazioni.

Un esempio di interrogazione che effettua una ricerca su tre diverse edizioni della Classificazione delle Professioni e sulla Classificazione delle attività economiche (Ateco) è rappresentato da una query JSON che esegue una richiesta GET di questo tipo:

```
GET cp2011,cp1971,cp1981,ateco/_search
{
    "from" : 0,
    "size" : 10,
    "query": {
      "match_phrase_prefix": {
          "nome" : {
            "query" : "elettrauto",
            "slop": 3
        }
      }
    },
    "highlight": {
        "fields" : {
            "nome" : {}
        }
    }
}
```

Il risultato è un output in formato JSON in cui sono elencati i primi dieci risultati estratti sulla base del punteggio (*score*) assegnato in funzione del potere discriminante della stringa cercata all'interno dei documenti è riportato in appendice.

Molte considerazioni affrontate in questo paragrafo devono essere necessariamente affiancate da alcune evidenze, che riguardano l'interattività tra i sistemi di ricerca e gli utenti. Il completamento automatico dei termini, ad esempio, pur essendo molto pratico, nel caso della ricerca diretta di un oggetto all'interno delle tassonomie induce facilmente l'utente ad una scelta frettolosa e imprecisa; lo stesso strumento, utilizzato all'interno di applicazioni ad hoc che fanno uso di dati collegati alle tassonomie, si è rivelato, invece, di estrema utilità. Anche la visualizzazione dei risultati ha un ruolo importante: cercare un elemento all'interno di una lunga lista articolata su più livelli può creare molta confusione. Quindi, nonostante in molti casi sia sensato effettuare delle codifiche a livelli di dettaglio più generici dell'ultimo, per numerose classificazioni statistiche è più utile fornire un output che contenga le istanze associate all'ultimo nodo e gli esempi ad esso correlati raggruppati al primo nodo.

Questa scelta è condizionata da una duplice motivazione: la ricchezza di termini nelle istanze di dettaglio maggiore e la facilità di indirizzare l'utente verso una ricerca corretta a partire dalla scelta del singolo nodo che per definizione è più generico e definisce le proprietà delle istanze sottostanti. Di solito, l'ultimo livello di una classificazione, corredato dagli esempi, contiene un dizionario molto ampio che permette di effettuare ricerche complete e precise. Tuttavia, ridurre la ricerca a una lista che contenga esclusivamente l'ultimo nodo potrebbe indurre gli utenti a una scelta errata a causa della confusione che introducono collezioni numerose di oggetti disomogenei.
Prendendo in esame ancora una volta la Classificazione delle Professioni, si consideri la ricerca del termine "meccanico". Una lista ordinata che comprende le istanze associate agli ultimi nodi e gli esempi ad essi correlati è qualcosa di simile:

	V° Livello
2.2.1.1.1	*Ingegneri meccanici*
3.1.3.1.0	*Tecnici meccanici*
6.2.2.2.0	*Costruttori di utensili, modellatori e tracciatori meccanici*
6.2.2.3.2	*Aggiustatori meccanici*
6.2.3.1.1	*Meccanici motoristi e riparatori di veicoli a motore*
6.2.3.1.3	*Meccanici di biciclette e veicoli assimilati*
6.2.3.2.0	*Meccanici, riparatori e manutentori di aerei*
6.2.3.6.0	*Meccanici collaudatori*
6.2.3.8.2	*Meccanici e motoristi navali*
6.2.4.1.3	*Elettromeccanici*
6.3.1.1.0	*Meccanici di precisione*
6.3.1.2.0	*Meccanici e riparatori di protesi, di ortesi, di tutori ortopedici e assimilati*
7.2.6.2.0	*Addetti a telai meccanici e a macchinari per la tessitura e la maglieria*

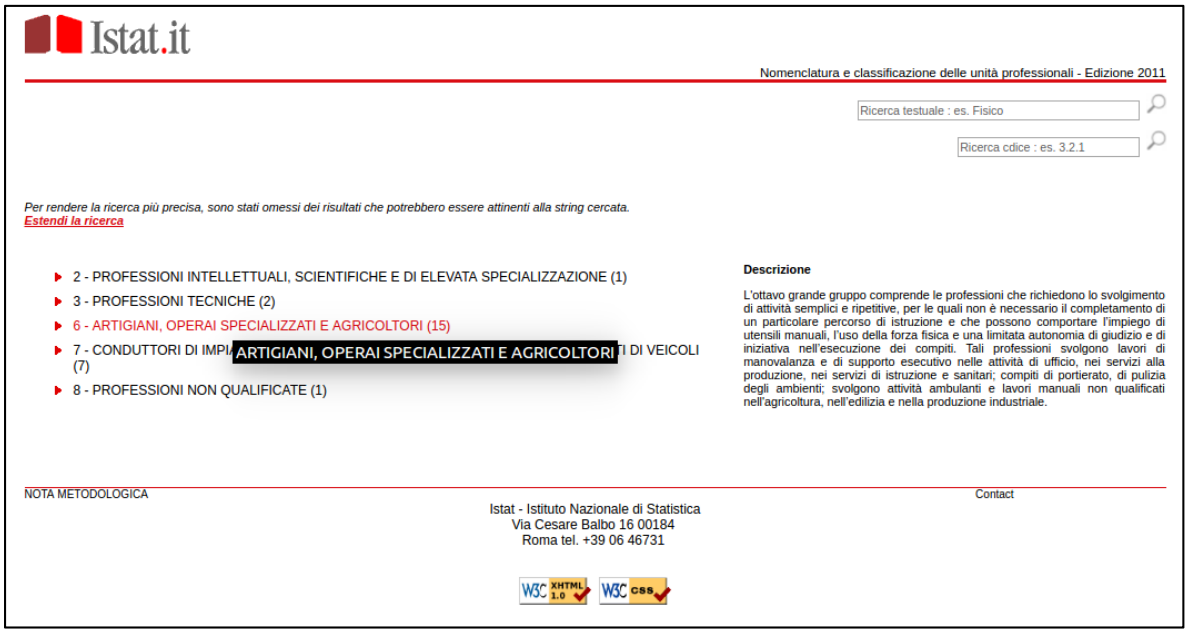
Esempi

6.2.2.3.2.1 aggiustatore meccanico di utensili
4.1.2.2.0.3 codificatore dati meccanografici
1.2.2.2.0.5 direttore o dirigente generale di azienda di costruzioni meccaniche
3.1.3.7.1.13 disegnatore meccanico
6.2.2.2.0.1 disegnatore tracciatore di sala (meccanico)
3.1.3.3.0.1 elettromeccanico di precisione di impianti nucleari
1.2.1.2.0.4 imprenditore o amministratore delegato di grande azienda di costruzioni meccaniche
1.3.1.2.0.3 imprenditore o responsabile di piccola azienda di costruzioni meccaniche
2.2.1.3.0.2 ingegnere elettromeccanico
2.2.1.1.1.1 ingegnere meccanico
6.2.3.2.0.2 meccanico aeronautico
6.2.3.1.1.2 meccanico di macchine agricole
6.2.3.2.0.3 meccanico di motori a reazione
6.2.3.1.1.3 meccanico di motori a scoppio
6.2.3.1.1.4 meccanico di motori diesel
6.2.2.3.1.16 meccanico fresatore
6.2.3.4.1.2 meccanico frigorista industriale
6.2.3.1.1.5 meccanico motorista
6.2.3.1.1.6 meccanico riparatore d'auto
6.2.3.3.1.2 meccanico riparatore di macchine a vapore
6.2.3.1.1.7 meccanico riparatore di motocicli
6.2.1.8.2.6 meccanico stampatore
6.2.2.3.1.17 meccanico stozzatore
4.1.2.2.0.10 operatore meccanografico
3.1.3.1.0.1 perito meccanico
6.2.2.3.2.4 piallatore meccanico
6.2.2.3.2.5 puntatore meccanico
3.1.3.1.0.2 tecnico calcolatore meccanico
3.1.3.1.0.3 tecnico conduttore di processo meccanico
3.1.3.1.0.4 tecnico di apparecchiature meccaniche di impianti nucleari

La difficoltà ad effettuare una codifica corretta, considerando i diversi gruppi a cui fa riferimento la parola meccanico, è evidente. Al contrario, obbligando l'utente a scegliere il gruppo del quale visualizzare i risultati, mostrandone una descrizione completa e articolata per evitare scelte errate, è possibile indirizzare più facilmente la ricerca verso un risultato preciso. Le figure seguenti mostrano un output organizzato come descritto, composto da due schermate:

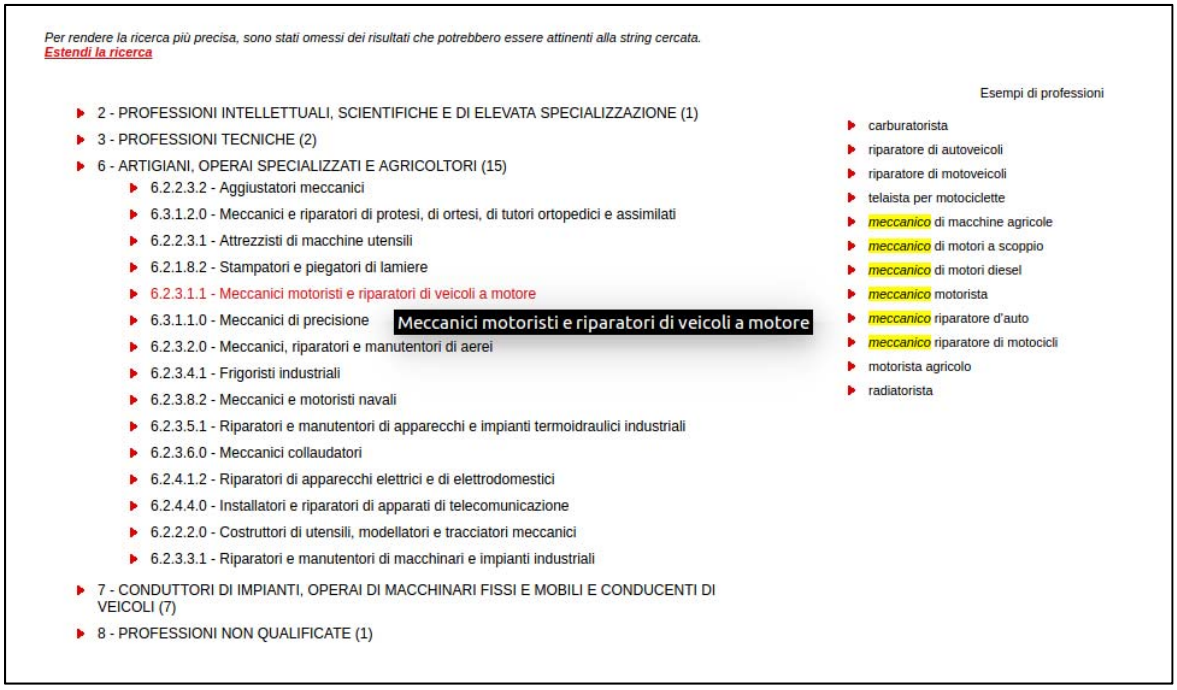
a) scelta del grande gruppo (Figura 2)
b) scelta del livello utile alla codifica (Figura 3)

Figura 2 - Output aggregato rispetto al nodo radice



Fonte: Elaborazioni realizzate attraverso l'uso del software per la ricerca semantica

Figura 3 - Output dettagliato relativo alle unità professionali

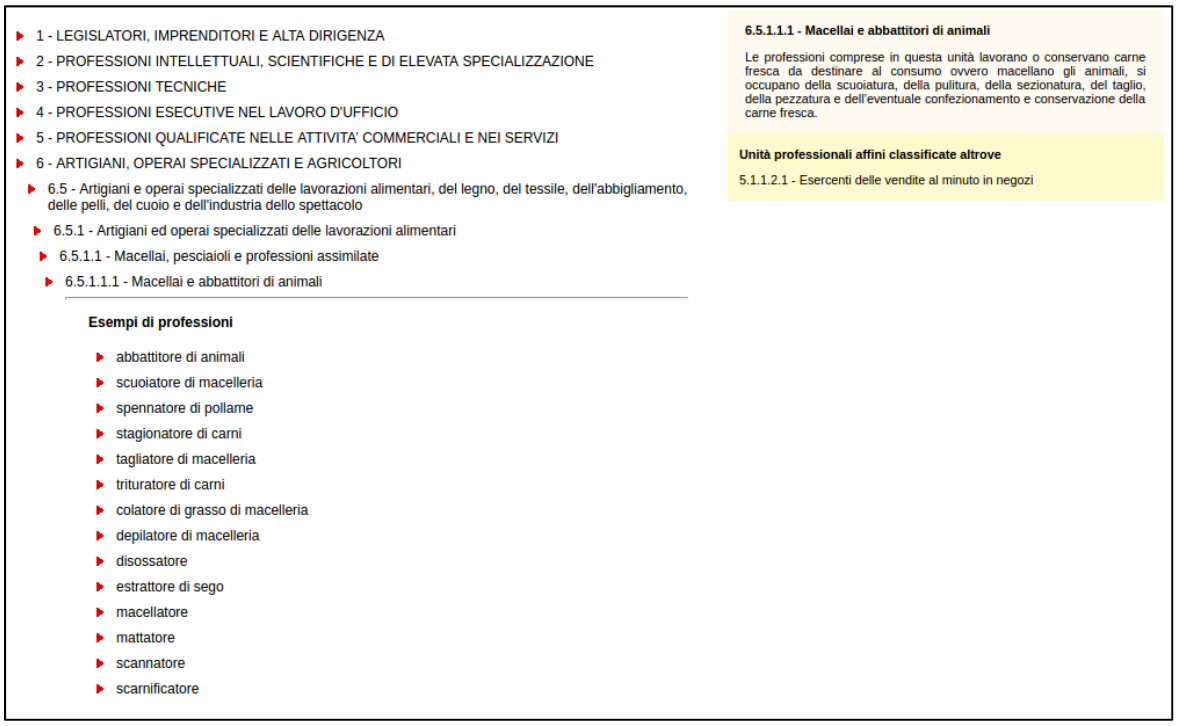


Fonte: Elaborazioni realizzate attraverso l'uso del software per la ricerca semantica

Per ridurre al minimo le possibilità di errori legati ai falsi risultati positivi, viene utilizzato un criterio di visualizzazione che prevede un primo output, all'interno del quale viene analizzato il punteggio (*score*) per restituire un elenco di risultati quanto più attinente alla stringa cercata, e un secondo output, ottenibile facendo clic su un link dedicato, che permette di estendere la ricerca a un set di risultati più ampio. L'esplorazione dei livelli intermedi, collegati al risultato della ricerca, si

ottiene a partire dalla scelta dell'ultimo. In questo modo, l'utente viene ricondotto a una scheda riassuntiva all'interno della quale è visualizzato l'intero albero decisionale corredato da eventuali rimandi o livelli affini. La Figura 4 si riferisce alla scheda relativa all'unità professionale "6.5.1.1.1 - Macellai e abbattitori di animali", che contiene il collegamento agli "Esercenti delle vendite al minuto in negozi". In questo modo, viene suggerita agli utenti la possibilità di ulteriori codifiche rispetto a quella proposta dal sistema di ricerca.

Figura 4 - Albero decisionale che sintetizza le caratteristiche di un'unità professionale



Fonte: Elaborazioni realizzate attraverso l'uso del software per la ricerca semantica

3.3 Albero decisionale

L'albero decisionale è un sistema molto raffinato di ricerca a cui si può ricorrere per trovare una collocazione per quei casi che rientrano nei falsi risultati negativi.
Gli alberi di decisione costituiscono il modo più semplice di classificare degli "oggetti" in un numero finito di classi. Essi vengono costruiti suddividendo ripetutamente i record in sottoinsiemi omogenei rispetto alla variabile risposta. La suddivisione produce una gerarchia ad albero, dove i sottoinsiemi (di record) vengono chiamati nodi e, quelli finali, foglie. In particolare, i nodi sono etichettati con il nome degli attributi, gli archi (i rami dell'albero) sono etichettati con i possibili valori dell'attributo (Dulli, Furini, Peron, 2009, p. 63).

È possibile rappresentare uno degli aspetti critici, legato alla ricerca nelle classificazioni, prendendo ancora una volta in esame la Classificazione delle Professioni. La ricerca di "account manager", una delle tante professioni non contemplata nel dizionario della CP2011, non produce, nella ricerca testuale, alcun risultato. Tuttavia, deve essere sempre possibile codificare i falsi risultati negativi in virtù del principio di esaustività delle tassonomie.
Per risolvere questo problema, oltre al classico diagramma ad albero costituito da nodi, rami e foglie derivanti dalla struttura gerarchica della classificazione (Figura 4), si può ricorrere alla creazione di un albero decisionale probabilistico, ovvero un sistema che, attraverso la misura dell'entropia associata ad una sequenza di domande (*training set*), permette di codificare un oggetto generalmente con pochi passi.

La costruzione di un sistema simile prevede necessariamente la formulazione di un insieme di domande che descrivano sufficientemente tutte le foglie dell'albero. Nel caso della Classificazione

delle Professioni, è utile pensare alle unità professionali come alle foglie dell'albero da descrivere. Il questionario sottopone all'utente una serie di quesiti volti a orientare la ricerca sulla base di risposte binarie del tipo SI/NO. Per quel che concerne le professioni, si può pensare a domande quali: "È una professione che si svolge in laboratorio?", "È una professione che richiede l'uso di utensili da cucina?", "Si svolge all'aperto?", etc. È utile evidenziare che il *training set* è tanto più efficace quanto più le domande proposte si avvicinano alla professione desiderata. La soluzione migliore, quindi, è l'applicazione di un algoritmo che sia in grado di scegliere la domanda n-esima sulla base delle risposte precedenti.

3.1.1 Teoria dell'informazione e algoritmo C4.5

Uno degli algoritmi più efficaci, introdotto da John Ross Quinlan nel 1993, è il C4.5 e prevede la misura dell'entropia per la costruzione e la potatura dell'albero decisionale. L'entropia dell'informazione misura l'ordine dello spazio degli elementi facenti parte dell'albero decisionale: un valore elevato di entropia rappresenta la difficoltà nell'assegnare ciascun elemento alla classe di appartenenza.

L'entropia di Shannon **H** di una variabile discreta **X,** che assume valori discreti $\{x_1 , x_2 \dots x_n\}$ le cui probabilità sono $P=\{p_1 , p_2 \dots p_n\}$, è definita come:

$$H(P)=-\sum_{i=1}^{n} p_i \times \log(p_i) \qquad (1)$$

In altre parole, se ci sono *n* messaggi possibili ognuno dei quali ha probabilità $p=\frac{1}{n}$, allora il valore $H(x)=log(n)$ rappresenta l'informazione contenuta nel messaggio.

Se **T** è il *training set* costituito da elementi preclassificati in **C_k** classi, si possono definire le probabilità $P=\{p_1 , p_2 , \dots , p_n\}$ come

$$P=(\frac{C1}{T},\frac{C2}{T},\dots,\frac{CK}{T}) \qquad (2)$$

L'informazione associata al *training set* **T** può essere definita come:

$$Informazione(T)=H(T)Informazione(T)=H(T) \qquad (3)$$

Se si effettua uno *split*, ovvero una segmentazione, di T in base al valore di un attributo X, supponendolo per semplicità discreto e categorico, si ottiene un nuovo partizionamento di T in $(T_1 , T_2 , \dots , T_n)$; a questo punto, l'informazione necessaria per identificare la classe di un elemento di T è:

$$Informazione(X,T) = \sum_{1}^{n} \frac{|Ti| \cdot Informazione(Ti)}{|Ti|} \qquad (4)$$

Si può quindi definire il guadagno di informazione (*Gain*) relativo a uno *split*:

$$Gain(X,T)=Informazione(T)-Informazione(X,T) \qquad (5)$$

Questo stimatore tende a favorire gli attributi che sono particolarmente numerosi, di conseguenza è meglio utilizzare un *GainRatio* definito in questo modo:

$$GainRatio(X,T)=Gain(X,T)SplitInfo(X,T) \quad (6)$$

dove lo *SplitInfo(X,T)* rappresenta l'informazione persa durante lo *split* sul valore discreto X. Supponendo uno *split* in $T=(T_1 , T_2 , \dots , T_n)$ dovuto al valore di X, lo *SplitInfo* viene calcolato come l'entropia di questa distribuzione utilizzando l'equazione definita sopra.

L'algoritmo C4.5 è definito da:

```
MakeTree (Training Set T)
                    Partition (T);
End  MakeTree

Partition (Set S)
                    if (all points in S are in the same class) then return;
                    Evaluate splits for each attribute A;
                    Use best split found to partition S into S1 and S2;
                    Partition (S1);
                    Partition (S2);
End  Partition
```

Si può dettagliare l'algoritmo C4.5 nei seguenti punti:

1) Il partizionamento si arresta quando non si identifica una segmentazione che possa ridurre significativamente la diversità di un dato nodo che diverrà un nodo foglia dell'albero. Nel momento in cui restano soltanto nodi foglia, l'albero sarà completo e l'algoritmo si arresterà.

2) Il criterio di segmentazione (*split*) della procedura *Partition* vale solo per attributi numerici o categorici. Per trovare la migliore partizione di un attributo, è necessario distinguerne il tipo (numerico o categorico).

L'albero costruito nella prima fase è utile per classificare dei *training set*, ma può causare un problema di *overfitting* (ipermodellamento o sovra-adattamento). Per ottenere un modello generale applicabile efficacemente, occorre potare l'albero, ovvero rimuovere i rami che possono condurre a errori nella classificazione dei dati, mantenendo il sotto-albero che presenta il minimo errore stimato. Per valutare l'errore associato a un albero, occorre riferirsi all'errore associato a ogni foglia (e_i):

$$e_i = n_i / n_{it} \qquad (7)$$

dove:
- n_i numero di dati classificati correttamente dalla foglia i,

- n_{it} numero totale dei dati presenti nella foglia i.

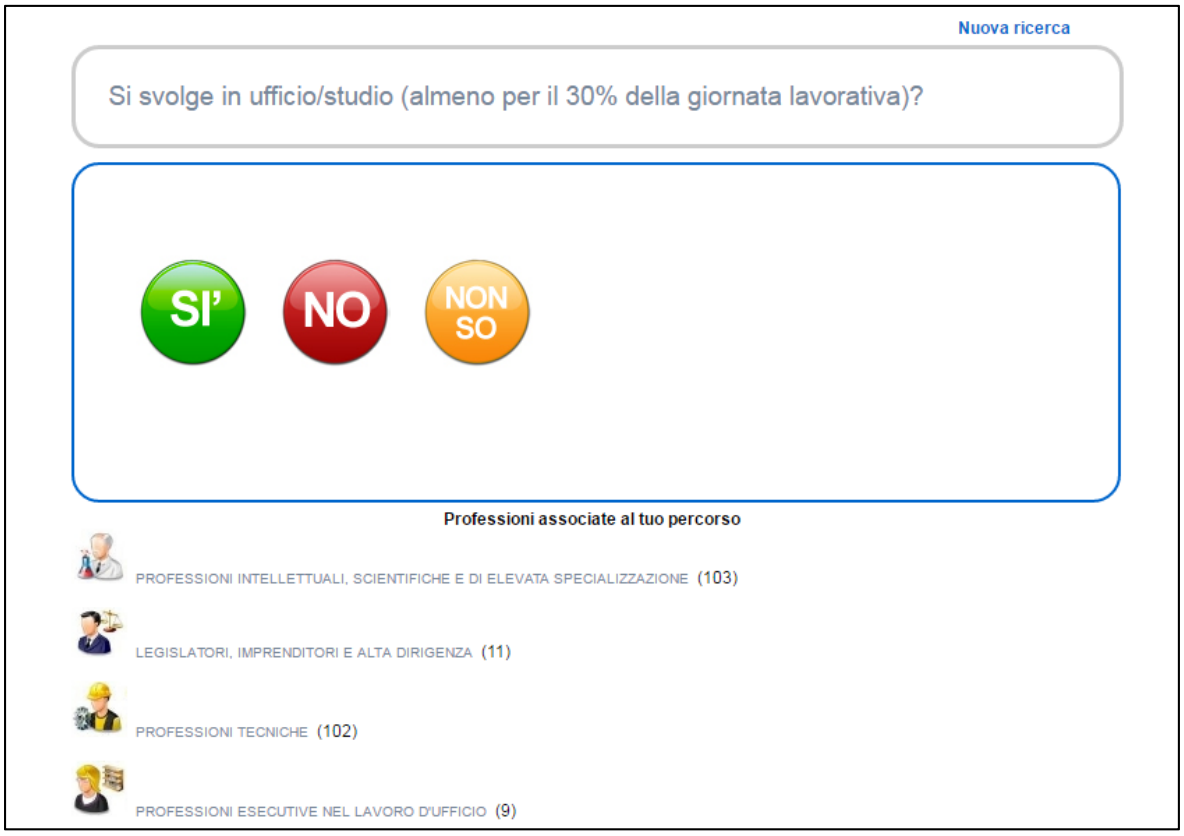
L'errore associato a tutto l'albero è invece la somma degli errori di tutte le foglie, pesati però con la probabilità che un nuovo record sia associato a ciascuna foglia:

$$E = \sum i\,(n_{it} / N)\, e_i \qquad (i = 1...n) \qquad (8)$$

Sulla base delle considerazioni sopra illustrate, è stato definito un *training set* di 125 domande trasversali per le 800 unità professionali. A ciascuna domanda è possibile fornire soltanto una risposta di tipo binario (Si/No). La risposta "Non so" consente di sottoporre all'utente una diversa domanda che abbia lo stesso contenuto informativo della precedente. L'utente segue un percorso, fornendo al sistema una sequenza binaria utilizzata per (Figura 5):

- Definire la domanda successiva
- Verificare la condizione di *split*
- Effettuare la potatura dell'albero
- Proporre un set di unità professionali sulla base della potatura

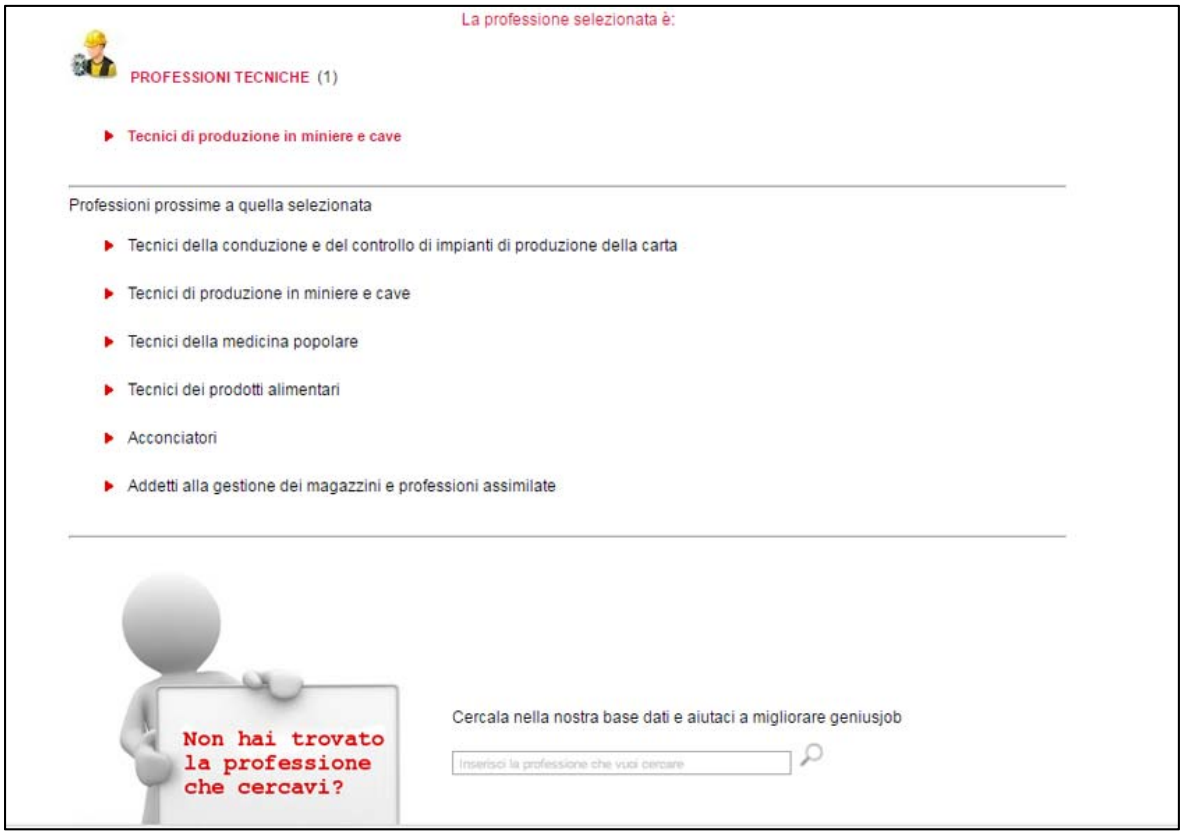
Figura 5 - Domanda n-esima e albero potato



Fonte: Elaborazioni realizzate attraverso l'uso del software per la ricerca semantica

Terminate le domande, il sistema fornisce una o più unità professionali che hanno la minore distanza di Hamming dal profilo contenuto nella sequenza binaria. Qualora l'unità professionale individuata non corrisponda alle aspettative dell'utente, viene proposto un modulo per segnalare, attraverso una ricerca testuale, la foglia corretta (Figura 6). L'algoritmo, infatti, è in grado di apprendere dalle segnalazioni degli utenti, per fornire ricerche più raffinate.

Figura 6 - Output dei risultati



Fonte: Elaborazioni realizzate attraverso l'uso del software per la ricerca semantica

4. Sistemi di codifica ad hoc

4.1 I dati collegati alle classificazioni

Le classificazioni ufficiali in alcuni casi possono rappresentare la struttura di base attraverso la quale delimitare il campo di un'indagine. È il caso dell'Indagine campionaria sulle professioni (realizzata da Istat e Isfol) che utilizza la CP2011 per rilevare 400 descrittori per tracciare il profilo delle unità professionali, di numerose indagini sulle imprese, che fanno largo uso dell'Ateco 2007, o dell'indagine Excelsior in cui vengono usate entrambe le classificazioni. In alcuni di questi casi, le variabili rilevate attraverso l'indagine possono fornire un valido supporto per la creazione di sistemi di ricerca ad hoc. La ricerca testuale, concettualmente, esaurisce le possibilità di codifica attraverso l'individuazione di un termine che corrisponde più o meno al "pensiero" del codificatore.

Molte tassonomie, però, utilizzano dei termini appartenenti ad un linguaggio distante dal linguaggio parlato. Come evidenziato nei paragrafi precedenti, ad esempio, la parola "impiegato", largamente utilizzata nel mercato del lavoro per circoscrivere un ampio numero di professioni, non ha lo stesso significato all'interno della Classificazione delle Professioni. Stessa criticità viene riscontrata nella classificazione ICD-10 e nell'Ateco. Questa distanza è dovuta al rigore scientifico con cui vengono costruite le tassonomie e in particolare al rispetto delle condizioni di esaustività ed esclusività. La corretta codifica è un punto cruciale di molte indagini (forze lavoro, dottori di ricerca, laureati, imprese, prezzi, cause di morte, etc.), alcune delle quali vengono svolte con tecniche CAWI (Computer Assisted Web Interviewing) e autocompilazione da parte degli utenti, che spesso non hanno nessuna conoscenza delle diverse logiche classificatorie.

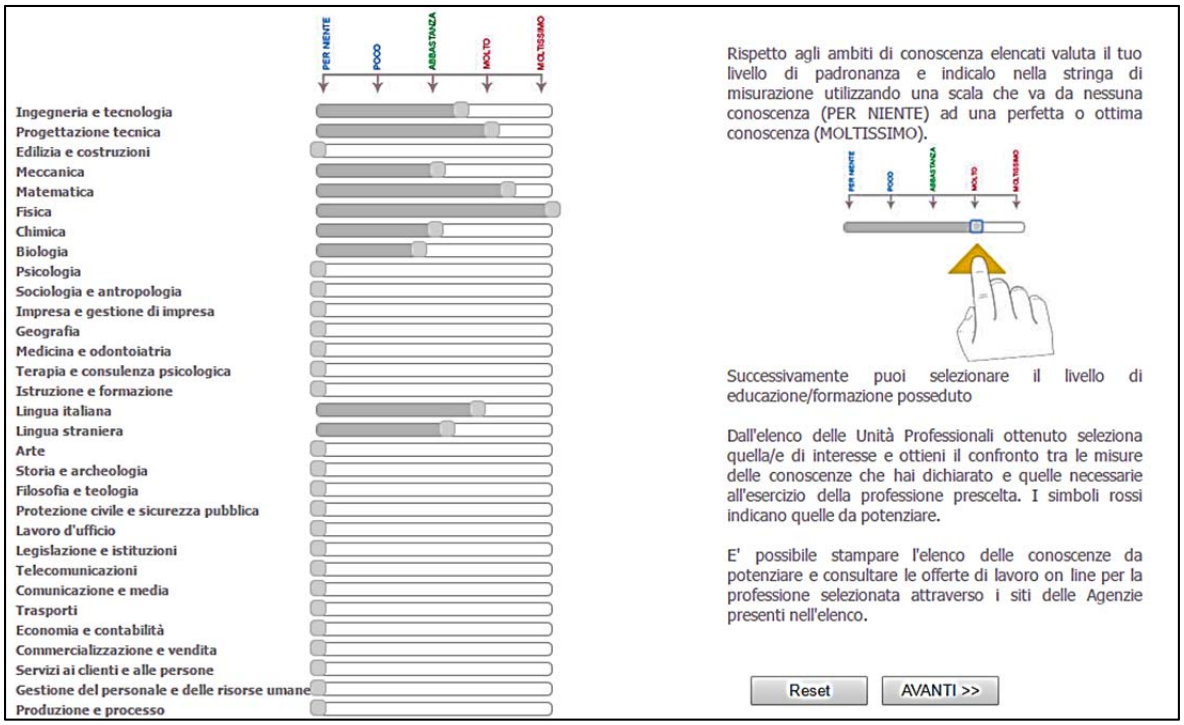
Per questo, è necessario oltrepassare i criteri con cui è costruita una tassonomia e fornire degli strumenti differenziati che facilitino l'individuazione del codice corretto. La Classificazione delle

Professioni e l'indagine ad essa associata si prestano molto bene a questo scopo e forniscono ampi spunti di riflessione per un'estensione di questa metodologia ad altri tipi di classificazione. In particolare, l'indagine rileva due aspetti molto interessanti legati alle unità professionali: le conoscenze che bisogna avere per svolgere una professione e i compiti svolti nell'ambito di una certa professione. Si tratta di due metodi di osservazione profondamente diversi in quanto le conoscenze vengono rilevate attraverso 33 variabili misurate in una scala da 0 a 100 sulla base dell'importanza e della complessità, mentre i compiti vengono rilevati attraverso una descrizione fornita dall'intervistato e standardizzata ex post rispetto alla totalità dei compiti rilevati. Normalmente, un utente che effettua una ricerca semantica pura all'interno di una classificazione, risponde alla domanda: "Qual è il nome dell'oggetto cercato?". Per rispondere a questa domanda, è necessario far ricorso ai termini conosciuti rispetto all'ambito circoscritto dalla classificazione utilizzata. I termini adoperati dagli utenti spesso non corrispondono a quelli contenuti nella tassonomia o assumono significati diversi che possono produrre risultati inaspettati e conseguenti codifiche errate. La rilevazione di ambiti tematici diversi connessi alle classificazioni permette di sottoporre a un utente domande diverse che permettono un approccio alternativo alla codifica.

4.2 Un sistema di codifica basato sulle conoscenze legate a una professione

Un sistema di ricerca per le professioni basato sulle conoscenze necessarie per esercitarle permette di affrontare il problema della codifica in un'ottica diversa, cambiando la domanda a cui rispondere, che in questo caso diventa: "Quali conoscenze servono per svolgere la professione che stai cercando?". A tale riguardo, la scelta migliore è sottoporre all'utente un questionario di autovalutazione delle conoscenze possedute, rispetto ai descrittori utilizzati nell'indagine. Il questionario è rappresentato nella Figura 7 sottostante.

Figura 7 - Autovalutazione delle conoscenze



Fonte: Elaborazioni realizzate attraverso l'uso del software per la ricerca semantica

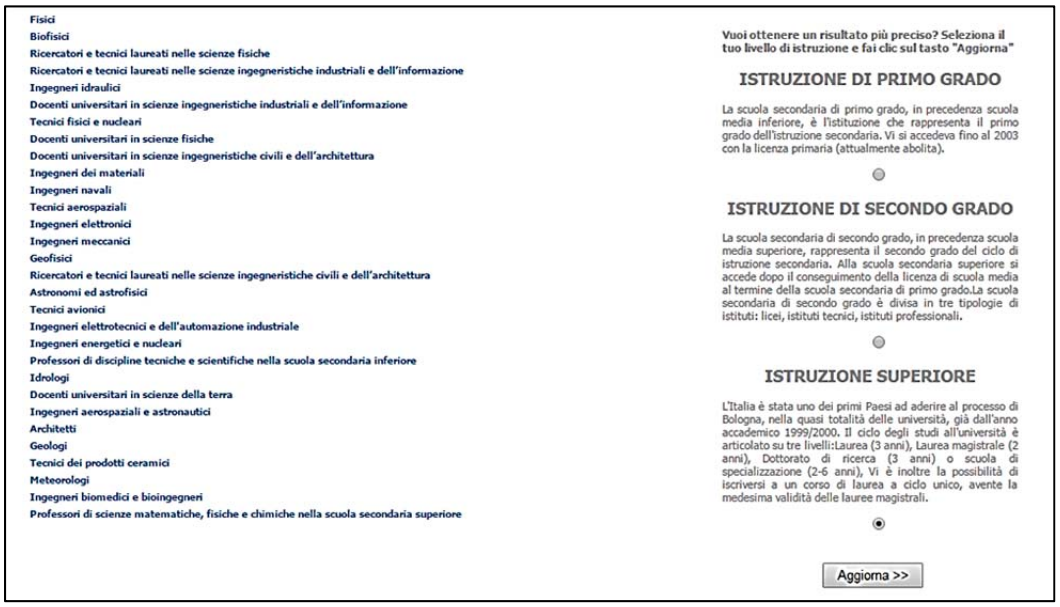
Le misurazioni degli ambiti di conoscenza fornite dall'utente possono essere confrontate con le misurazioni rilevate dall'indagine, per selezionare un set di unità professionali vicine a quella cercata. Poiché entrambe le serie di misure possono essere considerate come i valori assunti da due

variabili indipendenti, è possibile effettuare un confronto attraverso il calcolo dell'indice di correlazione *r* definito da:

$$r = \frac{\Sigma[(xi - X)(yi - Y)]}{\sqrt{[\Sigma(xi - X)2 - \Sigma(yi - Y)2]}} \qquad (9)$$

La Figura 8 riassume l'elenco delle unità professionali proposte dal sistema e ordinate sulla base dell'indice di correlazione rispetto alle misure rappresentate nella Figura 7.

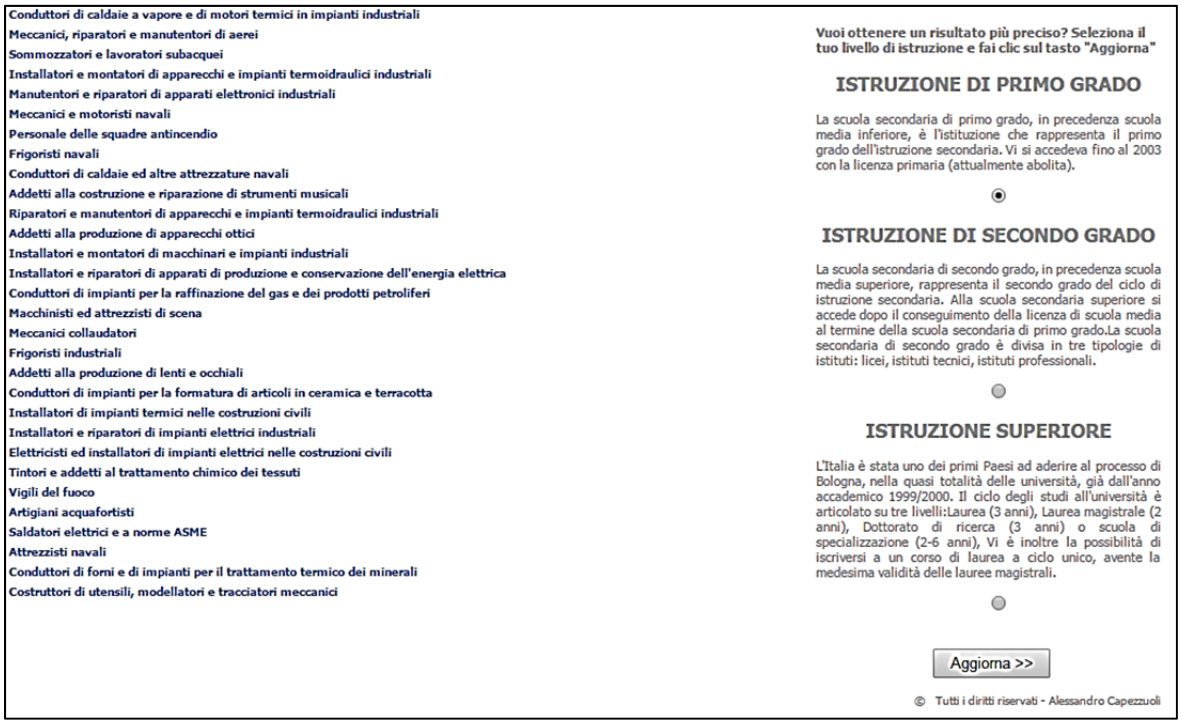
Figura 8 - Output delle unità professionali



Fonte: Elaborazioni realizzate attraverso l'uso del software per la ricerca semantica

Pur fornendo un output estremamente preciso, i risultati proposti dal sistema potrebbero non essere sufficienti a orientare l'utente verso una giusta codifica. Molte professioni, infatti, sono subordinate al titolo di studio posseduto: l'ulteriore scelta del livello di istruzione permette una selezione ancor più circostanziata (Figura 9).

Figura 9 - Output delle unità professionali filtrato in base al titolo di studio



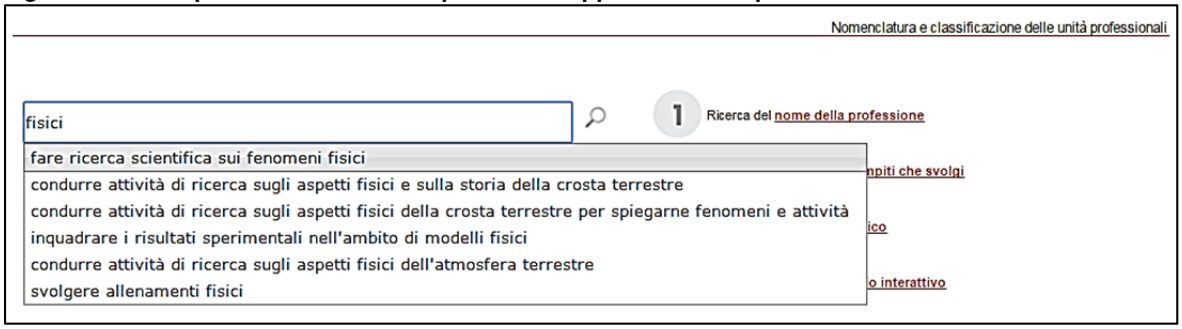
Fonte: Elaborazioni realizzate attraverso l'uso del software per la ricerca semantica

4.3 Un sistema di codifica basato sui compiti legati a una professione

La rilevazione dei compiti associati alle professioni fornisce un input interessante per la realizzazione di un ulteriore strumento di codifica che consenta di rispondere ad un altro tipo di domanda: "Qual è il compito principale che svolgi nell'ambito della tua professione?". In questo caso, le modalità di ricerca semantica necessitano di altre valutazioni rispetto alla ricerca dei termini della tassonomia. I compiti vengono rilevati utilizzando il linguaggio parlato; ciascun compito è rappresentato da una frase composta perlopiù da un soggetto, un predicato e un complemento. Il predicato individua e circostanzia molto bene un elenco di compiti, ma introduce un problema non trascurabile connesso ai diversi sinonimi associati.

La Figura 10 sintetizza alcuni dei compiti connessi alla professione "fisico". La frase "Fare ricerca scientifica sui fenomeni fisici" potrebbe essere cercata sotto una forma diversa e sintetica del tipo "Svolgere ricerca fisica". In questo caso, il tasso di insuccesso legato al *mismatch* potrebbe essere molto alto. Per risolvere i problemi associati alle innumerevoli possibilità offerte dalla lingua parlata, è necessario ricorrere ad alcuni strumenti che spesso non sono indicati per la ricerca diretta nelle tassonomie: il completamento automatico e i sinonimi.

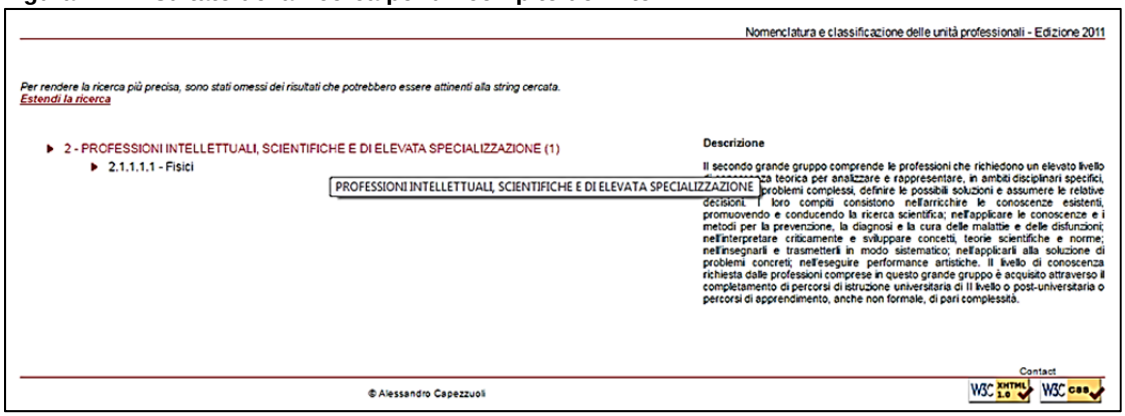
Figura 10 - Esempio di ricerca nei compiti con il supporto del completamento automatico



Fonte: Elaborazioni realizzate attraverso l'uso del software per la ricerca semantica

Il completamento automatico (Figura 10), che nella ricerca diretta all'interno delle tassonomie può indurre a scelte frettolose e imprecise, è invece indicato per suggerire a un utente delle frasi scritte in linguaggio parlato, che possono essere declinate in molti modi diversi. I sinonimi, uniti ai trattamenti di *parsing* della stringa, vengono invece utilizzati per favorire il *match* tra frasi che possono contenere parole simili. La costruzione del dizionario dei sinonimi prevede l'analisi testuale dell'universo osservato, il calcolo delle frequenze osservate e l'esclusione dei termini che potrebbero generare confusione: tanto più il dizionario dei sinonimi è confuso, tanto maggiore è la possibilità di ottenere falsi risultati positivi e codifiche errate. L'individuazione dei compiti relativi a una professione conduce molto spesso a una codifica esatta (Figura 11).

Figura 11 - Risultato della ricerca per un compito definito



Fonte: Elaborazioni realizzate attraverso l'uso del software per la ricerca semantica

È evidente che questo caso è ben diverso dall'approccio classico di ricerca diretta all'interno della tassonomia. L'utente è tenuto ad indicare un "concetto" estraneo alla struttura stessa, anche se fortemente collegato ad essa, ma probabilmente più familiare rispetto al linguaggio parlato.

4.4 Un sistema di codifica basato sull'indice della classificazione ICD-10

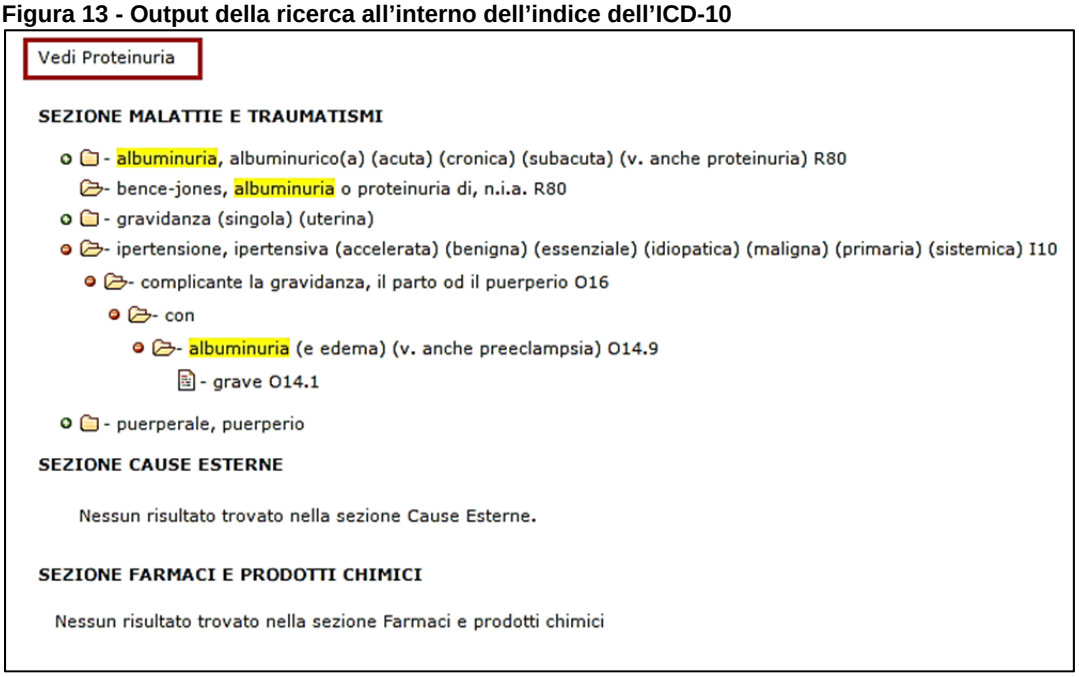
La classificazione ICD (International Classification of Diseases) ha una struttura gerarchica molto complessa. Per renderne più agevole la consultazione, l'Organizzazione Mondiale della Sanità ha corredato la tassonomia di un indice, a sua volta gerarchico, lessicalmente ricco e funzionale, che si presta molto bene per la creazione di un sistema di ricerca diverso da quello standard. L'indice è suddiviso in tre sezioni: Malattie e traumatismi, Cause esterne, Farmaci e prodotti chimici. Pur adottando ugualmente un approccio semantico, il sistema di codifica all'interno dell'indice presenta numerose peculiarità legate al dizionario utilizzato. Per illustrarne pienamente le caratteristiche, è necessario considerare alcuni esempi pratici. La ricerca del termine "Albuminuria", suggerito dal completamento automatico, fornisce l'output sintetizzato nella Figura 12.

Figura 12 - Interfaccia di ricerca all'interno dell'indice dell'ICD-10

Fonte: Elaborazioni realizzate attraverso l'uso del software per la ricerca semantica

I risultati della ricerca sono rappresentati in un albero gerarchico completamente diverso dall'albero della tassonomia, ma estremamente più rapido da consultare. Il collegamento con la classificazione è legato al codice identificativo della modalità, che consente un rimando preciso all'albero.

Figura 13 - Output della ricerca all'interno dell'indice dell'ICD-10

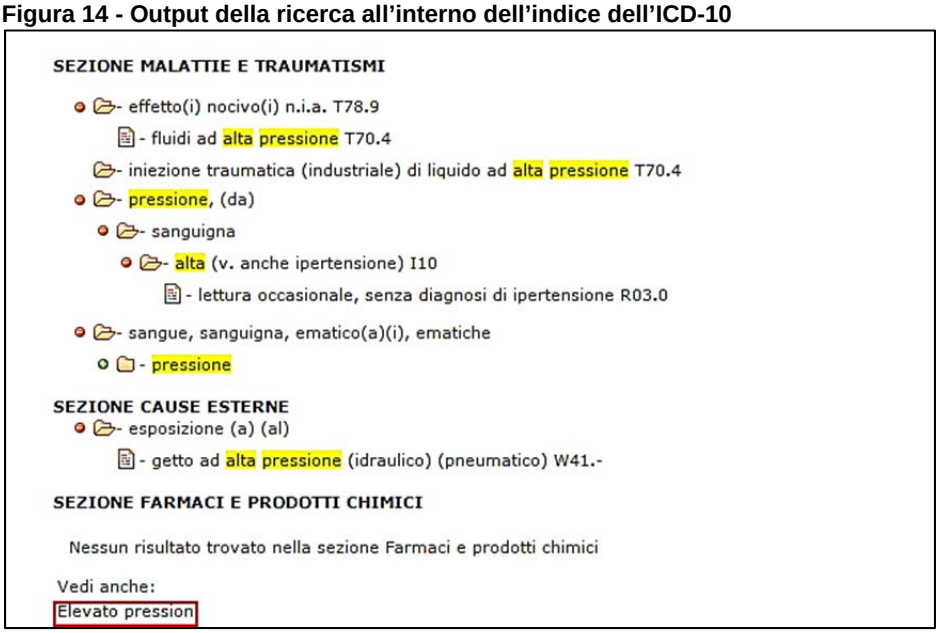


Fonte: Elaborazioni realizzate attraverso l'uso del software per la ricerca semantica

L'indice è corredato da un insieme di sinonimi (Figura 13, "Proteinuria") che permettono una ricerca precisa e puntuale dei termini non inclusi nel dizionario. Il lessico utilizzato contempla le diciture "Vedi" e "Vedi anche" per indirizzare l'utilizzatore alla modalità più appropriata.

Il suggerimento "Vedi" viene fornito all'utente per i casi in cui ad un termine corrisponda un falso risultato negativo, "Vedi anche" consente di estendere e raffinare la ricerca con termini più appropriati. Ad esempio, la ricerca delle parole "pressione alta" fornisce contemporaneamente dei falsi risultati positivi e dei falsi risultati negativi; questa anomalia è dovuta al fatto che all'interno dell'indice è presente il termine "elevata" al posto di "alta". Il sistema fornisce all'utente due tipi di indicazione in due step successivi. Il primo suggerimento riguarda la sostituzione della parola "alta" con "elevata". Pressione alta, infatti, oltre ad essere un termine utilizzato soltanto nell'indice e non all'interno dell'ICD-10, produce dei falsi positivi (fluidi ad alta pressione e getto ad alta pressione) in due sezioni distinte (Figura 14).

Figura 14 - Output della ricerca all'interno dell'indice dell'ICD-10



Fonte: Elaborazioni realizzate attraverso l'uso del software per la ricerca semantica

Facendo clic sul suggerimento, si sottopone al sistema una query più appropriata che restituisce dei risultati migliori (Figura 15) comprensivi dei falsi risultati negativi omessi dalla precedente ricerca.

Figura 15 - Output della ricerca all'interno dell'indice dell'ICD-10

Fonte: Elaborazioni realizzate attraverso l'uso del software per la ricerca semantica

È importante sottolineare che in questa fase l'utilizzatore viene guidato ad un'ulteriore ricerca più puntuale e precisa dal suggerimento "Vedi: ipertensione" che rappresenta il corretto termine medico da utilizzare all'interno della classificazione ICD-10 per avere dei risultati soddisfacenti.

5 Condivisione e disseminazione

5.1 Tecniche di condivisione e di disseminazione delle classificazioni

Non esiste un canale privilegiato per la diffusione e la disseminazione di una tassonomia e degli strumenti di ricerca/codifica ad essa collegati. Le esigenze possono essere fortemente diverse, di conseguenza i canali devono differenziarsi notevolmente per soddisfare ogni tipo di richiesta. Gli utilizzatori di una classificazione ufficiale sono molto eterogenei:

- Ricercatori
- Istituzioni pubbliche
- Cittadini
- Sviluppatori
- Aziende private

I canali per la condivisione sono altrettanto eterogenei:

- File di testo
- Web
- Web service
- Widget

Scegliere soltanto un canale di diffusione è sicuramente limitativo: gli utilizzatori possono avere esigenze molto diverse e avere necessità specifiche rispetto al contesto di applicazione. Tralasciando la diffusione attraverso il web, ampiamente analizzata nei paragrafi precedenti, e i file di testo, è opportuno focalizzare l'attenzione sugli strumenti di condivisione che consentono la cooperazione applicativa *machine-to-machine* e la consultazione dinamica. Molto spesso, è necessario includere i sistemi di codifica all'interno di altre applicazioni (siti web esterni, applicativi, questionari elettronici) per uniformare i criteri di consultazione e rendere disponibili i diversi algoritmi ai soggetti istituzionali che ne fanno richiesta.

I principali canali di diffusione sono i *web service* e i *widget* interattivi. I primi forniscono un output (SOAP o REST) strutturato in formato *machine readable* (XML/JSON) a seguito di un'interrogazione, i secondi sono delle vere e proprie applicazioni che forniscono un pacchetto per la codifica integrabile con lo stesso criterio con cui si incorporano i contenuti sui social network. Elasticsearch possiede un'architettura REST nativa che consente un'immediata diffusione dei contenuti indicizzati in formato JSON. Analogamente al contenuto degli indici, è possibile condividere gli algoritmi utilizzati per la ricerca, attraverso una query string parametrizzata. Questa soluzione, molto funzionale in termini di riuso delle classificazioni e di performance nei casi in cui gli utenti possano sviluppare dei software applicativi, è meno adeguata per integrare un vero e proprio sistema di codifica utilizzabile senza il vincolo dello sviluppo software. Un esempio pratico riguarda i questionari statistici nei quali sono presenti dei quesiti in cui viene richiesta una o più codifiche. In questo caso, la via più immediata è l'integrazione di un *widget*, ovvero uno script parametrizzato di questo tipo

<script
src="http://siprof.istat.it/siprof/classbrowser/widget.php?idIndice=1&idFamiglia=1&url=" type="text/javascript" ></script>

che consente, attraverso un semplice copia e incolla, l'inclusione di uno o più sistemi di codifica all'interno di qualsiasi applicazione esterna. I vantaggi di un *widget* parametrizzato che faccia riferimento ad un sistema centralizzato sono numerosi:

1) Al variare dei parametri *idIndice* e *idFamiglia* è possibile includere più classificazioni in diverse lingue.
2) Un *widget* realizzato con tecnologie javascript e web service REST consente comunque una cooperazione applicativa con i client.
3) Lo sviluppo di codice è molto limitato: basta includere poche righe per avere a disposizione un motore di ricerca potente e stabile.
4) La gestione degli aggiornamenti non ha alcun impatto sugli utilizzatori.

6 Conclusioni

La corretta codifica degli oggetti di una classificazione ha un ruolo fondamentale per la produzione di statistiche di qualità. La realizzazione di strumenti efficaci è essenziale per semplificare l'utilizzo delle tassonomie e standardizzare le metodologie di ricerca. Poiché le classificazioni hanno nature profondamente diverse, non esiste uno strumento o un algoritmo generalizzato che funzioni con successo, e con gli stessi criteri, su tutto. Le metodologie illustrate in questo documento, tuttavia, possono applicarsi alla maggior parte delle classificazioni, ufficiali e non. La tecnologia utilizzata, soprattutto nel caso della ricerca semantica, permette un alto livello di personalizzazione degli algoritmi attraverso la logica modulare parametrizzata. Essendo fortemente riferite a specifici campi di applicazione, le tassonomie possono essere più o meno collegate a dati aggiuntivi – per esempio provenienti dalle indagini o dagli archivi amministrativi – che consentono lo studio e la realizzazione di sistemi di codifica ad hoc, attraverso i quali oltrepassare la logica e i limiti della ricerca semantica. Per questo motivo, la progettazione di applicazioni per la codifica deve essere sempre contestualizzata rispetto alla tassonomia e all'ambito di conoscenza rappresentato. Un aspetto cruciale, infine, riguarda la diffusione delle classificazioni e dei sistemi di codifica; è necessario prevedere applicazioni web *mobile-friendly*, *web service*, *widget* e tutti quegli strumenti, molti dei quali illustrati in questo documento, che consentano di raggiungere facilmente utilizzatori differenti.

Appendice

Output Elsatisearch


```
{
  "took": 120,
  "timed_out": false,
  "_shards": {
    "total": 10,
    "successful": 10,
    "failed": 0
  },
  "hits": {
    "total": 4,
    "max_score": 7.7557683,
    "hits": [
      {
        "_index": "cp2011",
        "_type": "cp2011",
        "_id": "AVbWblmA2nUi1Tw46PcX",
        "_score": 7.7557683,
        "_source": {
          "pkLivello": "6.2.4.1.5",
          "nome": "Elettrauto",
          "notAnalyzed": "Elettrauto",
          "descrizione": "Le professioni comprese in questa unità installano, riparano e manutengono gli impianti e gli
apparati elettrici ed elettronici degli autoveicoli.",
          "fkLivello": "6.2.4.1"
        }
      },
      {
        "_index": "cp2011",
        "_type": "cp2011",
        "_id": "AVbWbln42nUi1Tw46PcZ",
        "_score": 7.7141705,
        "_source": {
          "pkLivello": "6.2.4.1.5.2",
          "nome": "elettrauto",
          "notAnalyzed": "elettrauto",
          "descrizione": "elettrauto",
          "fkLivello": "6.2.4.1.5"
        }
      },
      {
        "_index": "ateco",
        "_type": "ateco07",
        "_id": "AVbWw21i2nUi1Tw46c9n",
        "_score": 5.295668,
        "_source": {
          "pkLivello": "6124",
          "nome": "Elettrauto:  officina",
          "notAnalyzed": "Elettrauto:  officina",
          "descrizione": "Elettrauto:  officina",
          "fkLivello": "G.45.2.0.3.0"
        }
      },
      {
        "_index": "ateco",
        "_type": "ateco07",
        "_id": "AVbWw4mY2nUi1Tw46c_S",
        "_score": 5.283838,
```

```
        "_source": {
          "pkLivello": "6220",
          "nome": "Officine:  elettrauto",
          "notAnalyzed": "Officine:  elettrauto",
          "descrizione": "Officine:  elettrauto",
          "fkLivello": "G.45.2.0.3.0"
        }
        }
      ]
    }

}
```

Riferimenti bibliografici

APDOT - Advisory Panel for the Dictionary of Occupational Titles.1993. The New DOT: A Database of Occupational Titles for the Twenty-First Century. Washington, D.C.: U.S. Department of Labor, Employment and Training Administration.

Dulli S., Furini S., Peron E. 2009. Data mining Metodi e strategie. Springer.

ECIPA CNA, Fondazione G. Brodolini. 2003. Step Project. Stock and Trends in the European Professions. Donzelli.

Isfol. 2007. Nomenclatura e Classificazione delle Unità Professionali: Collana Temi&Strumenti-Studi e ricerche n. 36.

Marradi A. 1992. Concetti e metodo per la ricerca sociale. La Giuntina, Firenze.

Peterson N.G., Mumford M.D., Borman W.C., Jeanneret P.R., Fleishman E.A. 1995. Development of Prototype Occupational Information Network (O*NET) Content Model. Salt Lake City: Utah Department of Employment Security, 2 vol.

Informazioni per le autrici e per gli autori

La collana è aperta alle autrici e agli autori dell'Istat e del Sistema statistico nazionale e ad altri studiosi che abbiano partecipato ad attività promosse dall'Istat, dal Sistan, da altri Enti di ricerca e dalle Università (convegni, seminari, gruppi di lavoro, ecc.).

Coloro che desiderano pubblicare su questa collana devono sottoporre il proprio contributo al Comitato di redazione degli *Istat working papers*, inviandolo per posta elettronica all'indirizzo: iwp@istat.it.

Il saggio deve essere redatto seguendo gli standard editoriali previsti (disponibili sul sito dell'Istat), corredato di un sommario in Italiano e in Inglese e accompagnato da una dichiarazione di paternità dell'opera.

Per le autrici e gli autori dell'Istat, la sottomissione dei lavori deve essere accompagnata da un'e-mail della/del propria/o referente (Direttora/e, Responsabile di Servizio, etc.), che ne assicura la presa visione.

Per le autrici e gli autori degli altri Enti del Sistan la trasmissione avviene attraverso la/il responsabile dell'Ufficio di statistica, che ne prende visione. Per tutte le altre autrici e gli altri autori, esterni all'Istat e al Sistan, non è necessaria alcuna presa visione.

Per la stesura del testo occorre seguire le indicazioni presenti nel foglio di stile, con le citazioni e i riferimenti bibliografici redatti secondo il protocollo internazionale 'Autore-Data' del *Chicago Manual of Style*.

Attraverso il Comitato di redazione, tutti i lavori saranno sottoposti a un processo di valutazione doppio e anonimo che determinerà la significatività del lavoro per il progresso dell'attività statistica istituzionale.

La pubblicazione sarà disponibile su formato digitale e sarà consultabile on line gratuitamente.

Gli articoli pubblicati impegnano esclusivamente le autrici e gli autori e le opinioni espresse non implicano alcuna responsabilità da parte dell'Istat.