

Received 30 May 2022, accepted 18 June 2022, date of publication 27 June 2022, date of current version 5 July 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3186356

# A Semantic Approach to Ranking Techniques: Improving Web Page Searches for Educational Purposes

CARLA LIMONGELLI<sup>1</sup>, MATTEO LOMBARDI<sup>1</sup>, ALESSANDRO MARANI<sup>3</sup>, AND DAVIDE TAIBI<sup>1,4</sup>

<sup>1</sup>Department of Engineering, Roma Tre University, 00146 Rome, Italy

<sup>2</sup>Atos SE, 00142 Rome, Italy

<sup>3</sup>Compono Pty Ltd., Fortitude Valley, QLD 4006, Australia

<sup>4</sup>Institute for Educational Technology, National Research Council of Italy, 90146 Palermo, Italy

Corresponding author: Davide Taibi (davide.taibi@itd.cnr.it)

**ABSTRACT** The Web offers an unprecedented number of resources and has become the most popular source of information for students shaping their understanding of a new topic, and for instructors selecting relevant material for learning and teaching activities. Even though search engines are the most widely used tools for searching for educational content, the realities of the learning and teaching processes make the retrieval and evaluation of educational resources more complex than they are for other goods or services. The lack of recourse to educational metadata in web pages, as well as the size of the Web itself, call for specific techniques to be adopted for a more effective ranking of educational content. In this study, we propose an innovative approach based on semantic technologies. The *SemanticSearch* approach described in this paper leverages knowledge graph representation of teaching contexts and proposes a new ranking method for rating educational web content. In the literature we find an Educational Ranking Principle that ranks web pages for a specific teaching context. In this study, we integrate the Educational Ranking Principle with semantic data to extend the experimentation and analyse performance further. We undertake an evaluation involving university teachers, considering more than 70 queries to measure the *SemanticSearch* performance against the Educational Ranking Principle in addition to two state-of-the-art methodologies: Tf-Idf and BM25F. Paired t-tests of four accuracy measures provide statistical evidence for improvements made by using *SemanticSearch* method when compared to the three baselines.

**INDEX TERMS** Semantic-based retrieval, instructional materials, web ranking principles, teacher support.

## I. INTRODUCTION

The Web has become a primary source of information for a growing number of people, and it affects many aspects of our everyday life, education included. More and more frequently, the Web is the primary resource that students exploit to shape their understanding of a new topic. Teachers also take advantage of the myriad of resources available on the Web to create their lessons and to design learning activities. Other than web resources (courses, lessons) specifically created for educational purposes by numerous on-line educational providers, there is a vast number of web pages suitable for

The associate editor coordinating the review of this manuscript and approving it for publication was Hai Dong .

educational contexts, even though they have not been specifically designed for didactic purposes.

In this scenario, the retrieval of suitable resources that match a specific educational (or didactic) needs is crucial. As demonstrated by [1], retrieving and recommending educational resources cannot be directly compared to performing the same operations for other goods or services: information about the user's preferences or interests alone is not sufficient, because the data needed has to comply with educational requirements and learning objectives which go beyond the individual's profile. Such educational aspects must also be fully considered during the various phases of learning and teaching in order to arrive at a suitable recommendation of educational resources [2]. Moreover, it is important to

note that most education-oriented research into information retrieval and recommender systems considers students as the main users, while only a few studies focus on instructors [3] as the prime requesters of educational resources. Even in recent publications on emerging technologies for intelligent e-learning systems – as reported in the chapter “Technology in education” [4] – attention is mainly given to teaching methodologies (e.g., the “flipped classroom” approach) rather than to instructors’ problems in retrieving and reutilising teaching materials. Research into supporting learning and teaching design has been mostly oriented towards hybrid recommendation systems based on pedagogical patterns, metadata-based models for exploring materials, and intelligent systems to help the teacher save time when performing highly repetitive tasks [5]. However, teachers also need support (a) during the process of activity design, and (b) in the choice of relevant resources for learning and teaching [6]. In this scenario we envisage a web app for teachers that is able to support the implementation of our method for searching educational materials on the Web. To our knowledge, there are, as yet, no such apps, but only digital archives dedicated to educational content. An analysis of the problem of an effective retrieval of learning resources from an instructor’s perspective reveals specific requirements. The need to explore educational resources is probably at its peak when an instructor starts to teach a new course which is not related to previous classes; this may involve new topics or a different teaching context. Previously used resources might therefore not meet the instructor’s needs. Moreover, even when the same instructor teaches in different contexts, the resources delivered in one context may not match – or have little in common with – those needed in another [2], [6]. In general, we can say that when instructors are looking for new resources, it is very likely that their queries and requests will not conform to their previous teaching, either because the topic is new or because the educational aims of the course have changed.

The instructor’s educational context is critical in how information about their present teaching needs should be presented [2], [7]. We can deduce this context partly by looking at the structure of the course [8]. The course structure is, essentially, the representation of the knowledge and skills that the course aims to deliver. Such a structure, also known as the educational plan, can be visualised as the course concept map [9]. Moreover, an educational resource may be relevant when teaching a concept in just one specific context. For this reason, the retrieval of learning resources that match the educational context (contextual retrieval) is more important than suggesting resources based solely on an individual teacher or on past selections (user or content-based retrieval). For example, let us assume that a recommender system suggests a resource suited to the instructor’s teaching style, but offers a concept that is no longer part of the educational plan. It is clear that the recommendation is not wrong in itself since the resource is actually of interest to the instructor. However, it is rendered useless given the context of the current teaching

situation and so the recommendation is unsatisfactory from an educational point of view.

Our research focuses on the ranking of web pages according to their compliance with the educational context of the instructor. This focus on web pages introduces specific challenges regarding the following aspects:

- Metadata related to educational context are not widely adopted [10]. The use of Learning Resource Metadata Initiative LRMI increased over time but the distribution and quality of learning resources markup is limited. Therefore, while generic search engines leverage metadata to provide additional information matching a web search for products or services (e.g. prices for products), there are no educational annotation of resources.
- There are no standards applied for the structure of learning material published on the Web. Quality metadata to describe learning resources in dedicated learning object repositories are poor, and many incompatible standards are used for these metadata. Therefore, the majority of educators use common search engines when looking for educational resources [11].
- There is no specific educational rating for the content of Web pages. A mechanism to evaluate the capacity of using a content in an educational context does not exist. To support the search of learning materials, common search engines are still the predominant tools, even though they are not specifically oriented to educational content [11].

We thus propose a semantic search methodology that aims to give higher ranking to those web pages that are i) useful for the instructor and ii) concur with their current teaching requirements. This search methodology exploits information from the *Teaching Contexts* (TC) of the instructor as defined in a teacher model presented in [12]. That definition of TC includes the title, concept map, education level, teaching objectives and difficulty of the course. It also proposes a method called *Educational Ranking Principle* (ERP) that leverages the TC information to improve the retrieval of educational resources for a specific TC. Our new *SemanticSearch* method goes further, integrating semantic data extrapolated from the DBpedia knowledge graph with the TC information. In this way, it is no longer necessary to have suitable metadata content, dedicated standards or ratings for educational content to retrieve didactic material.

Knowledge graphs are widely adopted for organizing information about entities in a structured and semantically meaningful way. Essentially, when a teacher submits a query to retrieve didactic material from the Web, our *SemanticSearch* produces i) a Resource Category Graph and ii) a Query Category Graph. Our method combines the information from these two graphs and scores web pages according to how they fit in the elaboration of the two graphs. The higher the similarity score, the more likely the resource is suited to the teaching context of the educator.

Based on the information available in the TC, we tested our *SemanticSearch* methodology on 16 different queries,

created automatically as a combination of the components of the teaching context (as reported in Section III, Table 1). Queries are enriched through a Named Entity Recognition process with semantic entities from the DBpedia Knowledge Graph. Our methodology leverages specific properties and attributes of the semantic entities to improve ranking performance. The results of this semantic-based search are compared against two baselines: Tf-Idf, and BM25F. In our findings, the *SemanticSearch* is more reliable than Tf-Idf and BM25F in ranking resources according to their suitability in a teaching context, and it also improves on the already good results of ERP. In particular, some aspects of the TC in combination with semantic data help significant semantic entities to be detected more effectively, resulting in a better educational scoring of the web pages. In the following Section II we discuss related work in the field, while in Section III and Section IV we present the methodology and the evaluation of our *SemanticSearch* proposal respectively. Results are laid out in Section V, while Section VI presents final remarks.

## II. LITERATURE AND BACKGROUND

Our work investigates a ranking technique that can search for educational content throughout the entire Web. Discovering educational resources on the Web is a widely debated issue. In this respect, categorizing online content for didactic purposes is one of the main difficulties, mainly due to the lack of semantic relationships between Web resources that do not allow for effective automatic retrieval of teaching and learning material on the Internet. For instance, the absence of this semantic element makes it challenging to define relations between areas and subjects. Consequently, tasks such as finding associations between topics and listing recommendations about educational resources are still extremely difficult to perform automatically. Much research, particularly from the Technology Enhanced Learning (TEL) community, focuses on addressing these issues. In the following sections we will present some approaches that are closely related to our proposal.

An interesting work which exploits Linked Data available on the Web presents a process of supporting the semi-automatic classification of Open Educational Resources [13]. This method categories OERs based on the data enrichment capabilities that can be provided by social open vocabularies and archives. The authors demonstrate that it is possible to create a link between a formal knowledge organization system and a knowledge source. While ontology, a popular means of representing knowledge, can be used to categorise web pages with reference to conceptual schemes (e.g., thesauruses), a social categorisation using existing knowledge graphs, like DBpedia, can be more beneficial. Moreover, DBpedia resources are linked to ontologies such as YAGO and WordNet, thus providing more semantic information in the form of *typeOf* and *suchAs* relations.

The authors in [14] compare eight techniques for ranking semantic associations, highlighting two of them: size and entity homogeneity. The results prove that small semantic

associations between entities with similar types and semantic associations with uniform relations are, in practice, effective for human experts.

The importance of semantics is also highlighted in [15], where the authors demonstrate that using a semantic model to represent content is particularly effective in the retrieval of scientific articles. In [16], the authors present a novel solution for searching for and ranking OERs, which integrates the ranking of search results from an existing search engine with the rankings of OERs found via term clustering. The mixed ranking is then used to re-calculate the order of retrieved terms. Their goal is to improve the results of an existing search engine by increasing the number of relevant OERs relating to the search terms. A retrieval framework for vocabulary learning that is optimized for learning outcomes rather than general relevance is illustrated in [17]. The framework takes into account students' prior knowledge and optimizes the retrieval process by defining the learning and effort functions. Then, by combining them it is possible to obtain the final function for that optimization problem. In our case, we see the learning problem from a different point of view, namely, from the teacher's perspective. Our goal is to give stronger support to the teacher in arranging and retrieving the most appropriate learning material from the Web. In contrast to the approaches cited above, the focus of our research is the elicitation of educational content carried out by the actor of the search: the teacher. Our starting point is the *Educational Ranking Principle* (ERP) [12], an education-based ranking principle that elaborates the usefulness of a web page for a teaching context. The teaching context, included in the *Instructor Profile*, plays a fundamental role in the ranking. The formalisation of an instructor profile that is based on contextual information about the teaching is useful in addressing many problems when analysing web pages for educational purposes [18]–[21]. Leveraging the same information of the TC used for the ERP, studies in TEL have been able to address some important problems such as i) the automatic discovery of the prerequisites of educational resources [22], ii) potential new recommendation methods for instructors [23], [24], and iii) the comparison of the performance of recommender systems in TEL [25]. Recent research in this field also demonstrate that semantic entities of knowledge graphs are useful in enhancing the description of teaching resources [19], [26]. In this respect, our study also suggests a semantic-driven approach to further improve the scoring of web pages for a teaching context.

## III. METHODOLOGY

ERP aims to rate web pages according to specific aspects of an instructor's teaching practice. The rating reflects the matching of the web page with the concept to teach as well as the appropriate context. Google and other Information Retrieval (IR) methods already rank web pages according to topic and they perform remarkably. The main problem is ranking a set of web pages according to their suitability for teaching in a certain context; this is where ERP comes to

the fore. It is expected that if a web page is to be considered an educational resource, it not only needs to explain a concept, but also to refer to certain fundamental knowledge associated with it (e.g., prerequisite knowledge). It must also be appropriate for the target students. The Teaching Context is a key element in accurately ranking web pages according to the actual needs of the instructor [2], [12]. ERP is based on a specific structure of Teaching Context consisting of the following attributes:

- Concept Name (CN);
- Course Title (CT);
- Prerequisite Knowledge (PK);
- Difficulty (DIFF);
- Education Level (EL).

When teaching a concept, its name (CN), course title (CT) and prerequisite knowledge (PK) contextualise the scope of the teaching. A concept name may be too ambiguous, so a course title can be of help in describing the domain.

The prerequisite knowledge defines what the learners are expected to know already before acquiring the concept CN. Any teaching material that refers to its relative PK, as specified in the Teaching Context, will better suit the learning process. Let us suppose that an instructor is looking for educational resources for the concept  $c$  in a concept map. Let us also assume a simplified concept map, where the edges between concepts represent only one type of semantic relationship: the *prerequisite* relationship. The PK attribute is the set of concepts directly connected to  $c$  (i.e., with ongoing edges to  $c$ ). Prerequisite Knowledge has been valuable in addressing various problems in the retrieval and recommendation of resources in TEL [3], [22], [27], including the ranking of web pages [12]. Finally, the difficulty (DIFF) and the educational level (EL) of a concept tailor the search to those materials that the target audience can acquire appropriately.

These five attributes are sufficient in order to provide a higher score for those web pages that are most appropriate for the Teaching Context [12].

ERP analyses the content and structure of the web pages with respect to the Teaching Context. In brief, ERP is a structured scoring method that analyses the Tf-Idf score of terms of the teaching context in specific sections of the web pages. These scores are weighted according to the degree of expectancy in finding an attribute of the teaching context in a section of the web page. This mechanism is designed to assign a high score to web pages that present high Tf-Idf scores for an element of the teaching context in a significant section of the web page. When scoring a set of 10 web pages for a web search, the ERP scoring is more accurate and reliable than traditional IR methods such as Tf-Idf and BM25F. Performing the same test over a larger set of web pages (a more detailed account is given in Section IV-E), we discovered that ERP continues to perform better than the Tf-Idf baseline. The same improvement is also recorded when comparing ERP to BM25F, even though we were not able to obtain statistically significant results for this test. Over a large set of resources, the term frequencies suggest a good

match between a web page and the teaching context because of the high frequency of the terms in common. However, the resources may be very different on a conceptual level; this is where semantic techniques can assist IR methods [28], [29].

At this point we propose *SemanticSearch*, a revised ERP based on a semantic methodology rather than on pure term frequency. Using the same formulation of the teaching context, we introduce a potential semantic web search methodology to overcome the weaknesses of ERP. Figure 1 illustrates how *SemanticSearch* incorporates semantic techniques in the scoring phase of ERP. The main steps of the entire process are detailed in the following workflow:

**Step 0:** As a preliminary, the instructor wishes to create a course which comes with the following basic information: title (CT), educational level (EL), difficulty (DIFF) and a concept map which represents the organization of the course concepts based on their prerequisite relationships (PK).

**Step 1:** From the course information, the system generates the search query by combining the five educational attributes of the TC. In this study we aim to investigate which attributes of the TC lead to a better educational ranking of the web pages. In this regard, there are many possible query structures that can combine the attributes of the TC in different ways, as shown in Section IV-A1. In this study, we evaluate all of these and identify the one that is most relevant to our proposal. Once the system creates the query from the teaching context information, two different workflows emerge as Fig. 1 illustrates: (a) the creation of the semantic graph associated with the query, and (b) the creation of the semantic graph associated with the retrieved resources.

**Step a.1** Extraction of semantic entities from the query using the Dandelion Name Entity Recognition framework<sup>1</sup> [30].

**Step a.2** Creation of the semantic graph of the DBpedia categories associated with the query entities [31] by using the *dct:subject* property of each DBpedia entity extracted at Step a.1 (see Fig. 2).

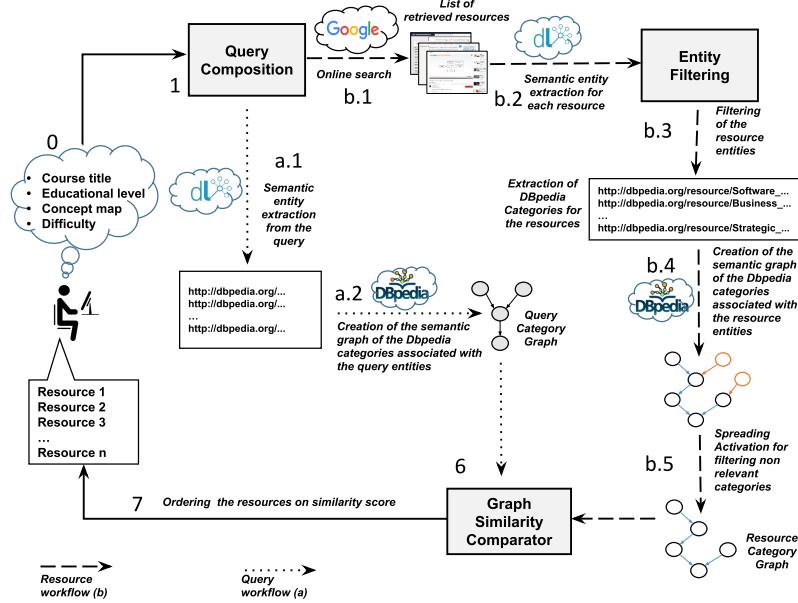
**Step b.1** Query submission. The query is sent to a search engine (e.g., Google) or ranking principle to retrieve the list of resources.

**Step b.2** Extraction of semantic entities from the content of each resource by using the Dandelion Name Entity Recognition framework. Semantically enriched entities extend the representation of entities from a mere sequence of terms to additional information about specific attributes. In our case the categories associated with each entity are leveraged for the ranking process.

**Step b.3** Filtering of the semantic entities extracted so that only the most promising entities are retained, ranked according to the specific scoring method (presented in Section IV-A2).

**Step b.4** Creation, for each resource, of the semantic graph from the DBpedia categories associated with the resource entities extracted during Step b.2.

<sup>1</sup><http://dandelion.eu/>



**FIGURE 1.** The seven steps of the operational workflow for a semantic ranking of web pages for education.

**Step b.5** In order to reduce noise and refine the overall category graph, we apply Dijkstra's algorithm and the Spreading Activation technique to the semantic graph created in the previous step. This step aims to remove categories that are likely to be irrelevant [31]. Specifically, after the extraction of the category tree from DBpedia, Dijkstra's algorithm is applied to find the shortest paths between the categories and the root of the category tree; for each edge belonging to these paths a weight equal to 1 is assigned; the spread activation algorithm is then applied to deduce the most important top-level categories for an entity. The activation is spread throughout the category graph contrary to the direction of the edges (from “child” to “parent”).

**Step 6** Computation of the similarity measure between the semantic graphs of the web pages and the semantic graph associated with the query. In this study we investigate which similarity measure, among those presented in Section IV-A3, is the most effective in ranking learning materials.

**Step 7** Ordering the resources according to the similarity score obtained from the Graph Similarity Comparator (during Step 6) by descending ratings.

#### IV. EVALUATION

The workflow in Figure 1 shows how the *SemanticSearch* methodology leverages semantic technologies for ranking instructional materials according to a teaching context. According to this methodology, the retrieval of web pages and the name entity recognition process rely on external services (Google Custom Search API and Dandelion NER respectively). However, i) the composition of the query structure, ii) the entity filtering and iii) the similarity of the query and resource graph are three key steps in the methodology that must be correctly tuned in order to improve ranking performance. In this work we evaluate several *SemanticSearch*

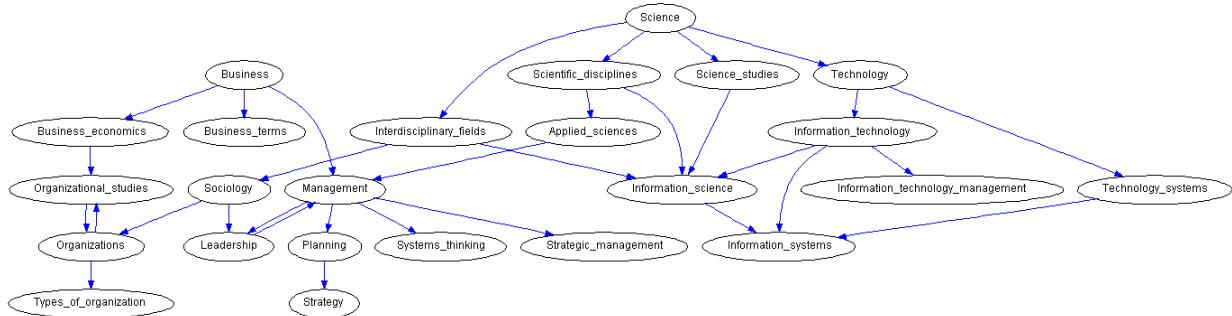
measures determined by the combination of the parameters involved in the three key elements of the workflow. The performance of these *SemanticSearch* measures is compared with the baselines Tf-Idf and BM25F using the Mean Average Precision (MAP) and top-N precision score as accuracy measures. The statistical t-test is used to confirm that the accuracy of our method is higher than the baselines. To produce a ground truth dataset of web pages rated within a teaching context, we asked 66 assessors to rate the relevance of web pages with respect to their teaching context. This dataset was used to compare the performances of our method against the baselines according to the MAP and top-N accuracy measures. In the following sections the elements of our evaluation methodology are presented in detail.

#### A. MEASURES

The *SemanticSearch* measures are obtained by combining the following three elements: Query structure, Scoring method and Similarity measure. The query structure is used in the *Query Composition* step to extract the attributes of the TC that will be used to generate the query to be submitted to the search engine. In the *Entity Filtering* step the scoring method is adopted to set the threshold to filter the entities that are associated with the resources (web pages) obtained from the search engine as results. Finally, the similarity measure intervenes in the *Graph Similarity Comparator* stage to determine the degree of similarity between the query and resource graphs. The following sections describe the three components in detail. Then, in Section V we evaluate which combination of these elements is the most beneficial for ranking the learning materials.

##### 1) QUERY STRUCTURES

When the system is asked to search for teaching material for a particular concept CN, the system can generate several



**FIGURE 2.** An example of a semantic graph produced from the categories of the entities extracted from the query “Business strategy, Strategic information system, Organizational strategy.”

**TABLE 1.** The set of query structures used for the evaluation of the *Semantic Search* methodology. The *Query Example* column shows an example of the query derived from the concept “Operators” for a “Java programming” undergraduate course.

Query id	Query Structure Attributes	#Attr.	Query Example
Q1	CN	1	Operators
Q2	CN, CT	2	Operators Java programming
Q3	CN, DIFF	2	Operators Beginners
Q4	CN, EL	2	Operators Undergraduate
Q5	CN, PK	2	Operators Variables
Q6	CN, CT, DIFF	3	Operators Java programming Beginners
Q7	CN, CT, EL	3	Operators Java programming Undergraduate
Q8	CN, CT, PK	3	Operators Java programming Variables
Q9	CN, DIFF, EL	3	Operators Beginners Undergraduate
Q10	CN, DIFF, PK	3	Operators Beginners Variables
Q11	CN, EL, PK	3	Operators Undergraduate Variables
Q12	CN, CT, DIFF, EL	4	Operators Java programming Beginners Undergraduate
Q13	CN, CT, DIFF, PK	4	Operators Java programming Beginners Variables
Q14	CN, DIFF, EL, PK	4	Operators Beginners Undergraduate Variables
Q15	CN, CT, EL, PK	4	Operators Java programming Undergraduate Variables
Q16	CN, CT, DIFF, EL, PK	5	Operators Java programming Beginners Undergraduate Variables

queries by combining the attributes of the TC. Since the order of the attributes does not influence the scoring phase, Table 1 reports the 16 query structures which represent all the combinations of the TC attributes. One aim of the present study is to identify which query structure, out of the 16 we can generate, is the best for the *Semantic Search* methodology.

Following the approach already presented in the literature [31], we produced a graph representation of the query using semantic entities extracted by means of the Dandelion API. In particular, the queries, written according to the query structures in Table 1, are elaborated with the Dandelion NER framework to obtain the corresponding DBpedia entities. Then, the *dct:subject* property indicates the related DBpedia categories. The hierarchical structure of the DBpedia categories is, in turn, used to construct the category graph of each entity.

For example, let us imagine that a teacher is searching for learning resources relating to the concept “Strategic

information system” (CN) for an undergraduate (EL) course titled “Business strategy” (CT) for beginners (DIFF). We also assume that the teacher specifies the concept “Organisational strategy” (PK) as a prerequisite for the Strategic information system. Figure 2 shows the category graph for the query “Business strategy, Strategic information system, Organizational strategy” automatically generated using the query structure Q8 in Table 1 with the attributes CN, CT, PK:

- Strategic information system (CN)
- Business strategy (CT)
- Organisational strategy (PK)

The Dandelion NER framework returns the DBpedia entities for this query: *dbr:Strategic\_information\_system*, *dbr:Strategic\_management*, *dbr:Organization*. The categories of these three entities are extracted using the *dct:subject* property of each entity. Successively, the category graph is created by considering the *skos:broad* relationships of the DBpedia categories. In the resulting graph, as shown in Figure 2, each node of the graph is a DBpedia category, and the top-level categories (Business and Science in the figure) have *dbc:Main\_topic\_classifications* category as *skos:broad*. The graph clearly shows that the user is asking for resources about strategic systems in the field of business. Generally speaking, the higher the number of terms in the query, the more complex the structure of the resulting query graph will be (i.e., with a higher number of nodes and edges). This is due to the fact that the entity extractor tool is likely to find a larger quantity of entities. This semantic representation of the query allows us to leverage semantic approaches in the overall ranking process.

## 2) SCORING METHODS

For each resource (web page) analysed in our method, we extract the semantic entities and then create a category graph following the same approach used for the query. The textual content of a resource may be much more detailed than a query, even just considering the number of words used. As previously stated, with increasing text length the graph structure displays a larger quantity of nodes and edges. However, the extractor may find entities that are poorly related to the resource as a whole. To reduce this noise and improve the performance of our method in the successive phases,

we introduce the filtering of the entities mined by Dandelion, thus retaining only the most relevant ones. We consider the following three possible approaches to filter the semantic entities:

- Ef-Irf: this refers to the calculation of an adapted version of the Tf-Idf (Term Frequency–Inverse Document Frequency) usually adopted in linguistic analysis. The Ef-Irf (Entity Frequency–Inverse Resource Frequency) measure considers entities rather than terms, and the resources to which the entities refer to rather than documents. We find this measure successfully enriches educational materials with semantic entities [32].
- Confidence: this is the value of confidence assigned by Dandelion to each extracted entity. The higher the confidence, the better the association mined by the tool. Thus, the entities extracted with greater confidence are likely to be the most significant ones.
- Ef-Irf \* Confidence: by multiplying the two scores obtained from the previous approaches we generate a third score that takes into account both frequency and semantic aspects.

### 3) SIMILARITY MEASURES

The final step is to produce a resource category graph out of the semantic entities extracted from a resource (web page) and filtered as mentioned above. The categories are linked to the entities through the *dct:subject* property, and, in addition, the categories are linked through the *skos:broader* property. Finally, each resource category graph is compared with the query category graph by the Graph Similarity Comparator. The similarity between two category graphs indicates if two documents, in our case the query and a resource (web page), address similar topics. There are different ways to compute the similarity between two category graphs. However, it is not the purpose of this study to investigate and compare such methods: here we are discussing just two of the approaches that we find most relevant to our study:

- Overlapping Degree (OD): this similarity measure addresses the different arrangement of common concepts between two graphs. It takes into account the perspective of the commonality of nodes, and the structural-semantic correspondence of the placement of the common nodes in the two graphs [33].
- Jaccard: this is a well-known measure that computes the intersection over union of the node sets of the two graphs. This measure does not consider the relations between nodes in the graph.

After this final step, resources are scored based on their relevance to the query. The semantic ranking is achieved by ordering these resources from the highest to the lowest score.

### B. BASELINE METHODS

The *SemanticSearch* methodology is an IR methodology. For this reason, its accuracy performance is compared to Tf-Idf and BM25F as they represent unstructured and structured

scoring methods respectively. We do not need to dwell on the importance of Tf-Idf and BM25F methods, which are the bases of modern IR approaches [34], [35] and are still used as the baseline for new IR methods [34]. These methods can be configured in different ways to reflect the specific scoring application better [34].

#### 1) TF-IDF

This measure is an unstructured scoring method that analyses the body text of a web page as a whole. The idea behind Tf-Idf is based on computing two coefficients for each term of a query, namely the Term Frequency (Tf) and the Inverse Document Frequency (Idf) in the collection of documents. The higher the Tf-Idf value for a document, the more relevant it is. Tf is the number of occurrences of a term in a document, while Idf is the number of documents that contain such a term in the dataset. When scoring documents for a non-binary query, a popular way of using Tf-Idf relies on a *Vector Space Model* (VSM) representation of both the query text and the document [35]. We thus build two vectors of Tf-Idf scores from the terms in the query: one for the body text of the web page and the other for the query text. The dimension of the vectors is equal to the number of unique terms in the query.

In this study, we use the Apache Lucene Practical Scoring Function<sup>2</sup> after removing the normalisation and boosting factors. Given a term  $t$  of a query  $q$  and the text  $text$ , we compute the Tf-Idf scores as follows:

$$\text{Tf-Idf}(t, text) = \sqrt{\text{frequency}(t \in text)} \cdot \text{idf}(t)^2, \quad (1)$$

where  $\text{idf}(t)$  is defined as follows:

$$\text{idf}(t) = 1 + \log \frac{\text{total number of documents}}{\text{docFreq}(t) + 1}.$$

Hence, the relevance score of a web page  $w$  for a query  $q$  is:

$$\text{Tf-Idf-Score}(q, w) = \text{cosine\_similarity}(V_q, V_w),$$

where  $V_q$  and  $V_w$  are the Tf-Idf vectors of the query  $q$  and the web page  $w$  respectively.

#### 2) BM25F

Unlike Tf-Idf, BM25F analyses the matching of a query with different parts or sections of a document. In the case of web pages, it is a straightforward process of extracting different kinds of information from the HTML tree of the pages. BM25F basically applies a BM25 function to score the relevance of diverse sections of a web page to a query. It then combines these scores using certain parameters to produce the relevance score. Section IV-E will provide an in-depth discussion of which parts of the web pages we consider for our purposes. However, it is useful to mention here that we consider these sections:  $s \in \{\text{title}, \text{body}, \text{links}, \text{highlights}\}$ . In our experiments, the setting of BM25F follows traditional methods [34], where we find that  $b_{\text{title}} = 0.4$ ,  $b_{\text{body}} = 0.3$ ,  $b_{\text{links}} = 0.4$ . For  $b_{\text{highlights}}$ , we have assigned a value of

<sup>2</sup>[http://lucene.apache.org/core/7\\_3\\_0/core/org/apache/lucene/search/similarities/TFIDFSimilarity.html](http://lucene.apache.org/core/7_3_0/core/org/apache/lucene/search/similarities/TFIDFSimilarity.html)

0.5 given that it is a section which is expected to represent fundamental concepts regarding the content of the web page. We set the value of  $K_1 = 1.7$  as per the original reference [34]. While the optimisation of boost factors may lead to better results for this method, we unfortunately do not have sufficient web searches for this purpose. Even the original reference with a much larger dataset could not optimise the parameters [34].

### C. ACCURACY MEASURES

When comparing the performance of IR methods, the main reference point is accuracy [36]. In this study, we analyse the *SemanticSearch* methodology following Mean Average Precision (MAP) and top-N precision scores.

#### 1) MEAN AVERAGE PRECISION

This is one of the most popular evaluation metrics in IR [36]. MAP makes it possible to compare the performance of two or more IR techniques over a set of queries. MAP expresses overall accuracy by computing the mean of the Average Precision (AP) scores of all the queries. Each query induces a subject method to rank the items. Given an order of items, the AP score is the average of the top-N precision scores in ranking the relevant items. In this study, we compare the MAP of our methodology against the baselines. To strengthen our findings further, we also comment on the paired t-tests of the AP scores relating to each query.

#### 2) TOP-N PRECISION

MAP already provides an analysis of the accuracy of the methods in terms of precision scores. In Information Retrieval, the Precision measure is computed as the fraction of the relevant documents from all those retrieved. However, further insight into the quality and practical benefit of a scoring method is provided by evaluating the Precision also considering the ranking of the relevant documents in an ordered results list. Let us consider a search engine that, for a query, shows all the resulting documents into pages, each containing 10 documents. Normally, a Web user tends to consider the web pages that appear on the first page and in the highest ranking positions as those that are most relevant. Therefore, when evaluating a ranking methodology it is important to take into account the position in rank for a relevant document: even when a relevant resource is correctly retrieved, the overall precision of the method is low if the methodology does not rank it in a top position. In contrast, a methodology that is able both to retrieve relevant resources and to rank them as top results yields high precision. For this reason, and to carry out a more practical appraisal of the methods, we compute three different values for the Precision measure in our evaluation, considering the relevant items in positions of high interest for the user: Precision at top-1 (denoted as  $P@1$ ), top-3 ( $P@3$ ) and top-5 ( $P@5$ ).

### D. STATISTICAL TESTING

The  $t$  distribution can be used for statistical hypothesis testing when only a small sample of the entire population has

participated in the study [37]. In such a scenario, the  $t$  distribution is found to be more reliable than the  $z$  distribution for hypothesis testing and constructing a confidence interval for the population mean [37].

The  $t$  distribution should be used when the underlying population is assumed to be normally distributed, and it is considered to be very accurate if the sample size is sufficiently large [37]. This applies to our case study.

The goal of our  $t$ -test is to show that our *SemanticSearch* methodology produces more accurate rankings than current practice. For each baseline approach, we perform a paired  $t$ -test with our proposal for the  $AP$ ,  $P@1$ ,  $P@3$  and  $P@5$  measures. In each test, the null hypothesis  $H_0$  states that the mean value of the accuracy of our method is lower than the baseline. Following our data quality assurance procedure, we recorded a total of 61 valid searches during the data collection phase. Given this sample size, we set the threshold for the  $t$ -value to 2.000 to reject the null hypothesis at .05 significance and 60 degree of freedom.

### E. DATASET

To run a comparison of the methods, we need a dataset of web pages with a relevance label for a teaching context. To the best of our knowledge, no such dataset exists, so we proceed with the collection of such data. During the data collection phase, we have to be sure that the external assessors rate the *usefulness* of web pages for teaching above their *relevance* to the query: a small but significant difference for a proper experiment [38]. For a reliable evaluation of the usefulness of a web page, external assessors rate items according to the highest level of knowledge and an awareness of the purpose of the web-search [38]. To fulfil these requirements, we designed and implemented an online survey that allows instructors to define a teaching context which is of interest to them, and which includes a concept map.<sup>3</sup> Users then formulate a query for retrieving web pages relating to a concept in their concept map. Our online survey system submits the query to Google,<sup>4</sup> and presents the assessors with the first 10 items as ranked by Google. We only select the top 10 items to keep the survey short, and because users tend to stop at the first page of items retrieved by a search engine (10 items in the case of Google with a standard configuration). To avoid any bias due to the system presentation order, the survey shows the 10 items in a random order, and assessors are made aware of this. The items are presented directly in a neutral environment so that the users do not know that Google is elaborating the query. Finally, the assessors rate the retrieved web pages according to their usefulness for teaching a concept in the teaching context. Assessors evaluate the web pages by means of a 5-point Likert scale, with each web page rated with a score from 1 to 5 (1 - not useful, 5 - extremely useful). Following

<sup>3</sup>It should be remembered that for our research, a concept map is a set of concepts taught in a course that are at least connected by the prerequisite relation.

<sup>4</sup>Google is queried by using the Google Custom Search service expanded to the entire web.

**TABLE 2.** HTML tags used for creating the structure of web pages content. For each of the four parts of a web page analysed, we indicate the HTML tags that are used for their composition.

Part of web page	HTML tags
Title	title
Body	body
Links	a
Highlights	strong, h3, h2, h1, b

this protocol, instructors are the external assessors of the web pages, and as such they define the contextual information, formulate the query, and rate the web pages. We believe that this data-collection protocol responds to issues highlighted by [38] in obtaining a true assessment of the *usefulness*, and not just the *relevance*, of items in a web search. After the data collection phase, the dataset includes 77 web searches, each of them with 10 web pages rated by the assessors.

To enable structured scoring of the web pages, we identify four main parts of a web page which are of interest for our application. Table 2 details the four sections that the methods can elaborate, and the HTML tags that we extract from a web page and the text of which makes up the four sections.

There is also the issue of eliminating text that is not relevant and that may generate some noise in the ranking task (e.g., menu-bars). This problem has no easy solution, and is beyond the scope of the present research. Usually, when proposing a ranking principle, such an issue can be resolved successively once the new approach has proved to work efficiently [34]. For our research, it is important that both the benchmark and our methods use the same texts during the experiments.

The presence of texts taken from noisy HTML tags also affects the scoring methods. We can thus prove the potential improvement of the new *SemanticSearch* measures in these conditions as well. We only perform some general cleaning of the texts (removal of stop words and stemming [39]) before we run any scoring methods. For all the queries and the corresponding pages retrieved by the search engine, we created a semantic representation with the corresponding query category graph and resource category graphs.

#### a: ADDITIONAL WEB PAGES TO RANK

In real-world scenarios, traditional IR scoring functions elaborate thousands of web pages in a web search, with only few of them being useful for the purpose of the web search. To make our evaluation similar to that scenario, we want to apply our scoring methods to a very large set of resources. However, it is not feasible to ask assessors to evaluate over 1,000 items for each web-search, and neither do we have the resources to collect such data from existing systems. Hence, on top of the 10 pages we collected during the data collection phase, we added at least 1,000 non-relevant web pages extracted from the DMOZ (now Curlie<sup>5</sup>) web directory. These pages are randomly selected from the different categories included in the directory. Considering that dead

links can be found in DMOZ, each search in the dataset may have a slightly different number of additional noisy items. On average, we added 1,070 noisy items, ending up with more than a thousand web pages to be ranked for the web search. Even though assessors have not rated the web pages from DMOZ, we know that the latter are not relevant as they come from non-educational categories of DMOZ (e.g., Shopping). We can assert that they are not useful for teaching in general (we do not make assumptions on their relevance to the topic of the query), so we label all the web pages from DMOZ with a rating of 1 (the lowest rating of usefulness).

For each DMOZ item that is added to the web-searches in our dataset, we create the semantic representation by i) extracting semantic entities, ii) filtering out irrelevant semantic entities, and iii) building the category graph according to the DBpedia categories of the entities.

## V. RESULTS

This section reports the outcome of the evaluation described in Section IV. The discussion compares the performances of our *SemanticSearch* and *ERP* measures against the baselines Tf-Idf and BM25F.

Since Tf-Idf and BM25F do not benefit from the educational information, while *ERP* and our *SemanticSearch* measures do, in order to make a fair evaluation we consider the accuracy of Tf-Idf and BM25F with both the queries formulated by the assessors and the queries built from the TC attributes (TC-based queries).

Because there are 16 possible TC-based queries (see Table 1), for the sake of brevity we compare our methods with the TC-based query that shows the best results for Tf-Idf and BM25F. The goal of this analysis is to identify the best combination of *query structure*, *scoring method* and *similarity* for our *SemanticSearch* methodology. We then compare the most accurate *SemanticSearch* measure with *ERP* and the baselines to appreciate the performance of our semantic methodology compared to existing practice. We will refer to each possible combination of the three components of the *SemanticSearch* using the following notation:

$<query\_id> <scoring\_method> <similarity\_measure>$ .

For instance, the Q1 query structure (see Table 1) with the Ef-Irf scoring method and the OD similarity measure is given the name *Q1EfIrfOD*.

However, we must consider that the query-graphs are automatically generated using the semantic entities taken from the query texts. For some of them, the Dandelion API could not mine any semantic entity even when the query contained academic or scientific terms. This issue is due to the nature of the queries that may not have enough terms that match with the DBpedia knowledge base. Since new semantic entities are continuously added in DBpedia, we expect this limitation of the proposed approach to have less impact in future versions of the knowledge base. To apply statistical analysis of our *SemanticSearch* measures against the baselines, we need to evaluate them over a set of at least 50 web searches. For this reason, we will only report the results of those methods that

<sup>5</sup><https://curlie.org/>

**TABLE 3.** Paired t-test analysis of the average precision of *ERP* against the baseline following the experimental methodology of this study.

Baseline	<i>t</i>	conf. diff.	mean	<i>p</i>
BM25F	1.435	-0.006	0.040	0.078
TC-based BM25F	1.070	-0.021	0.038	0.144
Tf-Idf	5.051	0.117	0.175	1.47E-06*
TC-based Tf-Idf	11.050	0.382	0.449	< 2.2E-16*

are able to produce the category graphs for at least 70% of the queries in our dataset. We decided on this threshold so each *SemanticSearch* measure could be applied to at least 54 web searches. This number is a compromise between having at least 50 web searches and enough *SemanticSearch* measures to compare.

The only *SemanticSearch* measures that were able to pass that threshold are: *q2EfIrfConfidenceOD*, *q5EfIrfOD*, *q8EfIrfJaccard*, *q10EfIrfOD*, and *q8EfIrfConfidenceJaccard*. Consequently, we evaluate here the performance of just these five measures.

Before going into the performance of the *SemanticSearch* measures, we first observe how the *ERP* performs against the baselines.

#### A. ERP VERSUS THE BASELINES

While we know that *ERP* works well with a small number of resources per search [12], performance drops significantly when we test the method over a large set of items per search. This is why we put forward the proposal for semantic approaches to be involved in the ranking process. Table 3 reports the paired t-tests of *ERP* against Tf-Idf and BM25F.

While *ERP* shows a strong improvement against Tf-Idf, BM25F is very challenging. This result confirms that a structured scoring of the web pages should be the focus of an educational ranking principle. Overall, *ERP* performs better than BM25F over the 77 searches of this study, but this is not enough to generalise the results on a large scale. We expect our *SemanticSearch* measures to achieve stronger results than *ERP*.

#### B. SemanticSearch VERSUS THE BASELINES

This section presents and discusses the accuracy performance of our *SemanticSearch* method against the accuracy of the baselines interrogated by the five queries that were able to pass the threshold illustrated above. Table 4 reports the results of the paired *t*-tests between the five *SemanticSearch* measures and the Tf-Idf. Since the *p*-values are lower than 0.05 for all of our methods, we conclude that *SemanticSearch* methods performed significantly better than Tf-Idf. However, *ERP* is already able to outperform Tf-Idf. The most challenging baseline is BM25F. When comparing the *SemanticSearch* methodology against BM25F, there is a milder outcome than with the previous results. Table 5 shows that, overall, only *q5EfIrfOD* performs better than the baseline, although there are some borderline situations. While *AP* and *P@5* leave no doubt as to the better ranking produced by the *q5EfIrfOD*

**TABLE 4.** The comparison of our *SemanticSearch* measures against the Tf-Idf baseline performed via paired T-tests. The "\*" symbol indicates a *p*-value lower than 0.05.

SemanticSearch	<i>t</i>	conf. diff.	mean	<i>p</i>
<i>q2EfIrfConfidenceOD</i>				
AP	5.410	0.140	0.204	0.0000004*
P@1	2.834	0.093	0.227	0.003*
P@3	4.694	0.127	0.197	0.000007*
P@5	5.655	0.117	0.167	1.881E-07*
<i>q5EfIrfOD</i>				
AP	<b>6.756</b>	<b>0.165</b>	<b>0.220</b>	<b>3.256E-09*</b>
P@1	<b>3.124</b>	<b>0.122</b>	<b>0.262</b>	<b>0.0014*</b>
P@3	<b>5.311</b>	<b>0.142</b>	<b>0.208</b>	<b>8.374E-07*</b>
P@5	<b>5.916</b>	<b>0.125</b>	<b>0.174</b>	<b>8.491E-08*</b>
<i>q10EfIrfOD</i>				
AP	6.010	0.149	0.207	5.92E-08*
P@1	2.847	0.102	0.246	0.003*
P@3	4.657	0.126	0.197	0.000009*
P@5	5.263	0.114	0.167	0.000001*
<i>q8EfIrfJaccard</i>				
AP	5.070	0.144	0.214	0.0000017*
P@1	1.797	0.011	0.149	0.0385*
P@3	4.773	0.155	0.239	0.0000052*
P@5	4.013	0.101	0.173	0.000078*
<i>q8EfIrfConfidenceJaccard</i>				
AP	5.665	0.142	0.202	1.741E-07*
P@1	2.257	0.047	0.179	0.014*
P@3	4.740	0.129	0.199	0.000006*
P@5	6.134	0.135	0.185	2.709E-08*

**TABLE 5.** The comparison of our *SemanticSearch* measures against the BM25F baseline performed via paired T-tests. The "\*" symbol indicates a *p*-value lower than 0.05.

SemanticSearch	<i>t</i>	conf. diff.	mean	<i>p</i>
<i>q2EfIrfConfidenceOD</i>				
AP	1.486	-0.008	0.062	0.071
P@1	0.554	-0.091	0.045	0.291
P@3	0.805	-0.038	0.035	0.2118
P@5	1.728	0.002	0.061	0.04437*
<i>q5EfIrfOD</i>				
AP	<b>2.269</b>	<b>0.022</b>	<b>0.085</b>	<b>0.01344*</b>
P@1	1.062	-0.056	0.098	0.1463
P@3	<b>1.750</b>	<b>0.003</b>	<b>0.071</b>	<b>0.0426*</b>
P@5	<b>2.285</b>	<b>0.021</b>	<b>0.079</b>	<b>0.013*</b>
<i>q10EfIrfOD</i>				
AP	1.788	0.005	0.071	0.03946*
P@1	0.869	-0.076	0.082	0.1943
P@3	1.352	-0.014	0.060	0.09071
P@5	1.303	-0.018	0.066	0.09853
<i>q8EfIrfJaccard</i>				
AP	1.528	-0.007	0.073	0.06562
P@1	-0.314	-0.188	-0.030	0.6228
P@3	1.413	-0.014	0.080	0.08122
P@5	0.870	-0.038	0.042	0.194
<i>q8EfIrfConfidenceJaccard</i>				
AP	1.586	-0.003	0.061	0.059
P@1	0.000	-0.123	0.000	0.5
P@3	0.915	-0.033	0.040	0.182
P@5	2.197	0.019	0.078	0.016*

method ( $p < 0.05$ ,  $t > 2.000$ , conf. diff.  $> 0.02$ ), *P@1* and *P@3* are more critical. We record an increase of 11.861% of *P@3* when we compare *q5EfIrfOD* to the average *P@3* value of the BM25F. While the *t*-test is not strong enough to support this result ( $t=1.75$ ), the *p*-value is lower than 0.05 suggesting a better *P@3* performance for *q5EfIrfOD*.

Overall, we find that the *q5EfIrfOD* approach is more reliable than traditional IR methods. Even though BM25F benefits from a formulation of queries, the statistical tests confirm that our *q5EfIrfOD* is still able to generate better rankings. Hence, we have been able to improve on *ERP* using semantic data in order to make it more robust when ranking an extremely large set of resources for a web search.

## VI. CONCLUSION AND FUTURE WORK

This study has addressed the critical issue of ranking educational material for educational applications. Web searches imply various challenges owing to the extremely large number of web pages involved, most of which are not relevant for the purposes of the search [40]. This situation is even more problematic when educational issues are involved. Not all web pages extracted by Google or other search engines are suitable for use in an educational context. Consequently, instructors are obliged to spend a lot of time and effort checking the many results from Google or other search engines before they can select the appropriate resources for their educational needs. An initial study of this topic suggests adopting *ERP*, which is a new structured method for scoring web pages according to teaching context [12]. However, this approach has some limitations when scoring a large number of web pages, i.e., not just the top 10 pages retrieved by Google.

To overcome this drawback, we have integrated *ERP* with semantic data. The proposed methodology combines Information Retrieval and Semantic Web methodologies, leveraging DBpedia knowledge graphs in describing the content of a query based on terms used in the instructor's teaching plan. While we find that *ERP* struggles when ranking hundreds of resources according to teaching context, the semantic data fix this issue. Proof of the validity of this statement is provided by the experimentation carried out. The dataset provides more than 1,000 noisy items in addition to the 10 Google answers rated by assessors as useful. Following the methodology illustrated in Section IV we established that the optimal composition of the query based on teaching context is the one that considers CN and PK. As a filtering method we used Ef-Irf, while as a measure of the similarity between graphs we used the overlapping degree. This combination (*q5EfIrfOD*) led us to establish that queries composed in this way not only perform very well in comparisons with the Tf-Idf baseline (Tab. 4), but also in comparisons with BM25F (Tab. 5). The paired *t*-tests regarding average precision against the baselines are positive. Moreover, the *P@5* and *P@3* metrics have a significant *p*-value (*p* < 0.05).

This study fulfills the promise of an educational ranking principle that is able to score a large number of web pages simultaneously in order to facilitate the retrieval of instructional materials. This is thanks to its semantic-based approach.

However, there are some limitations, with *ERP* and the *SemanticSearch* being useful in different applications. The term-based *ERP* is efficient and requires less data processing because it does not need to conduct a semantic analysis of

the teaching context and the web page content. On the other hand, these semantic data are crucial for ranking hundreds of web pages at once. The issue with semantic data extraction is not only computational, but also regards feasibility in certain contexts. During the study, we were able to apply the semantic approach to a reduced amount of web searches than the original *ERP*, namely 66 instead of 77 (86% of the total amount). The reason for this is that it was not possible to extract a sufficient number of entities for some teaching contexts. Therefore, *ERP* is a better application for increased web searches where a set of resources has already been pruned by another system, such as Google. In contrast, Semantic *ERP* works well on its own and does not necessarily have to be used in combination with another system. Since our methodology explains how semantic data can be used for the scoring of educational web pages, we expect a combination of textual and semantic analysis to provide a more scalable and accurate *ERP*. In conclusion, this study aims to minimize the time instructors have to spend looking for teaching material by automatically performing the time-consuming activity of evaluating the educational suitability of a web resource for a specific educational context. We also believe that current and future systems in TEL can benefit from *ERP* and *SemanticSearch* in order to exploit the huge amount of knowledge hosted on the Web efficiently. Such a development is likely to enhance the effectiveness of the overall teaching experience.

## REFERENCES

- [1] K. Verbert, N. Manouselis, X. Ochoa, M. Wolpers, H. Drachsler, I. Bosnic, and E. Duval, "Context-aware recommender systems for learning: A survey and future challenges," *IEEE Trans. Learn. Technol.*, vol. 5, no. 4, pp. 318–335, Oct./Dec. 2012.
- [2] J. Bozo, R. Alarcón, and S. Iribarra, "Recommending learning objects according to a teachers' context model," in *Sustaining TEL: From Innovation to Learning and Practice*. Berlin, Germany: Springer, 2010, pp. 470–475.
- [3] H. Drachsler, K. Verbert, O. C. Santos, and N. Manouselis, "Panorama of recommender systems to support learning," in *Recommender Systems Handbook*. New York, NY, USA: Springer, 2015, pp. 421–451.
- [4] M. Al-Emran, K. Shaalan, and A. E. Hassanien, *Recent Advances in Intelligent Systems and Smart Applications*. Cham, Switzerland: Springer, Jun. 2020.
- [5] O. Zawacki-Richter, V. I. Marín, M. Bond, and F. Gouverneur, "Systematic review of research on artificial intelligence applications in higher education—Where are the educators?" *Int. J. Educ. Technol. Higher Educ.*, vol. 16, no. 1, pp. 1–27, Dec. 2019.
- [6] K. Verbert, X. Ochoa, M. Derntl, M. Wolpers, A. Pardo, and E. Duval, "Semi-automatic assembly of learning resources," *Comput. Educ.*, vol. 59, no. 4, pp. 1257–1272, Dec. 2012.
- [7] M. Ferreira-Satler, F. P. Romero, V. H. Menendez-Dominguez, A. Zapata, and M. E. Prieto, "Fuzzy ontologies-based user profiles applied to enhance e-learning activities," *Soft Comput.*, vol. 16, no. 7, pp. 1129–1141, Jul. 2012.
- [8] V. Estivill-Castro, C. Limongelli, M. Lombardi, and A. Marani, "DAJEE: A dataset of joint educational entities for information retrieval in technology enhanced learning," in *Proc. 39th Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, Jul. 2016, pp. 681–684.
- [9] J. D. Novak and A. J. Cañas, "The theory underlying concept maps and how to construct and use them," in *Proc. IADIS Int. Conf. e-Learn.*, 2007, pp. 443–450.
- [10] S. Dietze, D. Taibi, R. Yu, P. Barker, and M. d'Aquin, "Analysing and improving embedded markup of learning resources on the web," in *Proc. 26th Int. Conf. World Wide Web Companion (WWW Companion)*. Geneva, Switzerland: International World Wide Web Conferences Steering Committee, 2017, pp. 283–292.

- [11] R. Cortinovis, A. Mikroyannidis, J. Domingue, P. Mulholland, and R. Farrow, "Supporting the discoverability of open educational resources," *Educ. Inf. Technol.*, vol. 24, no. 5, pp. 3129–3161, Sep. 2019.
- [12] V. Estivill-Castro and A. Marani, "Towards the ranking of web-pages for educational purposes," in *Proc. CSEDU*, 2019, pp. 47–54.
- [13] J. Chicaiza, N. Piedra, J. Lopez-Vargas, and E. Tovar-Caro, "Domain categorization of open educational resources based on linked data," in *Knowledge Engineering and the Semantic Web*, P. Klinov and D. Mouroumtsev, Eds. Cham, Switzerland: Springer, 2014, pp. 15–28.
- [14] G. Cheng, F. Shao, and Y. Qu, "An empirical evaluation of techniques for ranking semantic associations," *IEEE Trans. Knowl. Data Eng.*, vol. 29, no. 11, pp. 2388–2401, Nov. 2017.
- [15] D. Dessì, F. Osborne, D. R. Recupero, D. Buscaldi, and E. Motta, "Generating knowledge graphs by employing natural language processing and machine learning techniques within the scholarly domain," *Future Gener. Comput. Syst.*, vol. 116, pp. 253–264, Mar. 2021.
- [16] M. A. H. Pimentel, I. B. Sant'Anna, and M. D. D. Fabro, "Searching and ranking educational resources based on terms clustering," in *Proc. 20th Int. Conf. Enterprise Inf. Syst.*, Funchal, Portugal, Mar. 2018, pp. 507–516.
- [17] R. Syed and K. Collins-Thompson, "Retrieval algorithms optimized for human learning," in *Proc. 40th Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, Aug. 2017, pp. 555–564.
- [18] V. Estivill-Castro, M. Lombardi, and A. Marani, "Improving binary classification of web pages using an ensemble of feature selection algorithms," in *Proc. Australas. Comput. Sci. Week Multiconf.*, Jan. 2018, p. 17.
- [19] C. Limongelli, M. Lombardi, A. Marani, and D. Taibi, "Enrichment of the dataset of joint educational entities with the web of data," in *Proc. IEEE 17th Int. Conf. Adv. Learn. Technol. (ICALT)*, Jul. 2017, pp. 528–529.
- [20] S. Maloney, A. Moss, J. Keating, G. Kotsanas, and P. Morgan, "Sharing teaching and learning resources: Perceptions of a university's faculty members," *Med. Educ.*, vol. 47, no. 8, pp. 811–819, Aug. 2013.
- [21] K. Krieger, "Creating learning material from web resources," in *The Semantic Web. Latest Advances and New Domains*. Cham, Switzerland: Springer, 2015, pp. 721–730.
- [22] C. De Medio, F. Gasparetti, C. Limongelli, M. Lombardi, A. Marani, F. Sciarrone, and M. Temperini, "Discovering prerequisite relationships among learning objects: A coursera-driven approach," in *Proc. Int. Conf. Web-Based Learn.* Cham, Switzerland: Springer, 2016, pp. 261–265.
- [23] C. Limongelli, M. Lombardi, and A. Marani, "Towards the recommendation of resources in coursera," in *Intelligent Tutoring Systems: 13th International Conference, ITS 2016, Zagreb, Croatia, June 7–10, 2016*, vol. 9684. Cham, Switzerland: Springer, 2016, p. 461.
- [24] C. De Medio, F. Gasparetti, C. Limongelli, M. Lombardi, A. Marani, F. Sciarrone, and M. Temperini, "Towards a characterization of educational material: An analysis of coursera resources," in *Proc. Int. Symp. Emerg. Technol. Educ.* Cham, Switzerland: Springer, 2016, pp. 547–557.
- [25] M. Lombardi and A. Marani, "A comparative framework to evaluate recommender systems in technology enhanced learning: A case study," in *Advances in Artificial Intelligence and Its Applications*. Cham, Switzerland: Springer, 2015, pp. 155–170.
- [26] D. Dessì, G. Fenu, M. Marras, and D. R. Recupero, "COCO: Semantic-enriched collection of online courses at scale with experimental use cases," in *Proc. World Conf. Inf. Syst. Technol.* Cham, Switzerland: Springer, 2018, pp. 1386–1396.
- [27] C. Limongelli, A. Marani, F. Sciarrone, and M. Temperini, "Measuring the similarity of concept maps according to pedagogical criteria," *IEEE Access*, vol. 10, pp. 27655–27669, 2022.
- [28] Y. J. Du and Y. F. Hai, "Semantic ranking of web pages based on formal concept analysis," *J. Syst. Softw.*, vol. 86, no. 1, pp. 187–197, 2013.
- [29] V. Jindal, S. Bawa, and S. Batra, "A review of ranking approaches for semantic search on web," *Inf. Process. Manage.*, vol. 50, no. 2, pp. 416–425, Mar. 2014.
- [30] S. Parmesan, U. Scailla, M. Barbera, and T. Tarasova, "Dandelion: From raw data to dataGEMs for developers," in *Proc. CEUR Workshop*, vol. 1268, Jan. 2014, pp. 1–6.
- [31] C. Limongelli, M. Lombardi, A. Marani, and D. Taibi, "Enhancing categorization of learning resources in the dataset of joint educational entities," in *Proc. 16th Int. Semantic Web Conf. Posters Demonstrations Ind. Tracks (ISWC)*, 2017. [Online]. Available: <http://ceur-ws.org/Vol-1963/paper578.pdf>
- [32] D. Taibi, B. Fetahu, and S. Dietze, "Towards integration of web data into a coherent educational data graph," in *Proc. 22nd Int. Conf. World Wide Web*. New York, NY, USA: Association for Computing Machinery, May 2013, pp. 419–424.
- [33] C. Limongelli, M. Lombardi, A. Marani, F. Sciarrone, and M. Temperini, "Concept maps similarity measures for educational applications," in *Proc. Int. Conf. Intell. Tutoring Syst.* Cham, Switzerland: Springer, 2016, pp. 361–367.
- [34] J. R. Pérez-Agüera, J. Arroyo, J. Greenberg, J. P. Iglesias, and V. Fresno, "Using BM25F for semantic search," in *Proc. 3rd Int. Semantic Search Workshop*, 2010, pp. 1–8.
- [35] J. Ramos, "Using TF-IDF to determine word relevance in document queries," in *Proc. 1st Instruct. Conf. Mach. Learn.*, vol. 242, 2003, pp. 133–142.
- [36] G. Shani and A. Gunawardana, "Evaluating recommendation systems," in *Recommender Systems Handbook*. New York, NY, USA: Springer, 2011, pp. 257–297.
- [37] R. Witte and J. Witte, *Statistics*. Hoboken, NJ, USA: Wiley, 2013.
- [38] J. Mao, Y. Liu, K. Zhou, J.-Y. Nie, J. Song, M. Zhang, S. Ma, J. Sun, and H. Luo, "When does relevance mean usefulness and user satisfaction in web search?" in *Proc. 39th Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, New York, NY, USA, Jul. 2016, pp. 463–472.
- [39] M. F. Porter, "An algorithm for suffix stripping," *Program*, vol. 14, no. 3, pp. 130–137, 1980.
- [40] S. Brin and L. Page, "Reprint of: The anatomy of a large-scale hypertextual web search engine," *Comput. Netw.*, vol. 56, no. 18, pp. 3825–3833, 2012.



**CARLA LIMONGELLI** received the Ph.D. degree in computer science from the Sapienza University of Rome, Italy, in 1993. She is currently an Associate Professor with Roma Tre University, Rome. Her research interests include educational recommender systems, user modeling, and user-adapted interaction. She has organized some international workshops, congress tracks, and special issues in the area of artificial intelligence in education.



**MATTEO LOMBARDI** received the Ph.D. degree in computer science from Griffith University, Nathan, QLD, Australia, in 2019. Currently, he works as a Systems Architect with Atos SE, Rome, Italy. His research interests include web data analysis, natural language processing, application development on low-code and no-code platforms, and machine learning applied to education.



**ALESSANDRO MARANI** received the Ph.D. degree in computer science from Griffith University, Nathan, QLD, Australia, in 2018. He is currently a Data Scientist with Compono Pty Ltd., Fortitude Valley, QLD, Australia. His main research interests include data modeling and engineering, cloud computing, user profiling, and machine learning applied to education and HR industry.



**DAVIDE TAIBI** is a Senior Researcher with the Institute of Educational Technology, National Research Council of Italy; and a part-time Lecturer with the Department of Computer Science, University of Palermo. His main research interests include pedagogical applications to mobile learning, learning analytics, and social media in education. He is coordinating two European funded projects on the development of data literacy competencies for university and business sectors.