# Towards the Ranking of Web-pages for Educational Purposes

Vladimir Estivill-Castro and Alessandro Marani
*School of Information and Communication Technology, Griffith University, 170 Kessels Rd, Nathan QLD 4111, Australia*

Keywords: Educational Ranking Principle, Information Retrieval, Technology Enhanced Learning.

Abstract: The World-Wide-Web is a well-established source of resources for different applications and purposes including the support to learning and teaching tasks. The notion of Learning Object (LO) was specifically designed for sharing digital learning materials over web-applications enabling repositories of LOs. But, the extension of such repositories is rather small compared to the Web, and some of these repositories are domain-dependent. LOs typically provide some educational metadata describing the content. However, the WEB hosts hundreds of thousands of web-pages with educational content but with no educational metadata. Generic search engines provide the best current support to sieve such educational web-pages. But such present systems are not educational focused, so they may not pick instructional features that the users want or need for their educational task. We study a web-based retrieval method for using the Web as a repository of educational resources. Our proposal is a new structured scoring method named *Educational Ranking Principle* (*ERP*). *ERP* analyses the suitability of a web-page for teaching a concept in a specific educational context. Our approach shows a superior accuracy performance than Google, TFIDF and BM25F. The results of our experiment using MAP and P@1 undoubtedly confirm the improvement of *ERP* when compared to all the baselines (with a $p$-value less than 0.05). Moreover, *ERP* is the only method where our results have statistical support for higher accuracy than Google for all the four accuracy measures we use in this study.

## 1 INTRODUCTION

The Web is a well-established source of resources and services for different purposes including teaching and learning. The Web hosts a considerable number of web-pages where some of them have educational content. Google and YouTube are just two of the many platforms commonly used by instructors and learners for their educational tasks (Maloney et al., 2013). Also, a new format of online courses, called Massive Open Online Courses (MOOCS), enriches the Web with high-quality courses and materials (Kay et al., 2013; Limongelli et al., 2016a; De Medio et al., 2016a).

When using the Web as a dataset of documents for a specific purpose, the main issue is the design of an Information Retrieval (IR) method for the retrieval of useful documents (Brin and Page, 2012). In particular, ranking principles assess web-pages to produce ordered lists of items where useful items are placed in high positions. While many proposals of IR techniques and Recommender Systems (RS) in Technology Enhanced Learning (TEL) focus on Learning Objects (LO) stored in online repositories (Drachsler et al., 2015), the Web is usually left out. However, instructors and students rely more on Google and YouTube than Learning Object Repositories (LOR) (Maloney et al., 2013). We believe the main limitation of LORs is the number of resources they host compared to the Web. On the other hand, Learning Objects provides educational metadata of their content for enabling a more educational-oriented retrieval of the items. Given this diametric situation, we focus on how to expand the search and sieve learning resources from web-pages even though web-pages provide no educational metadata.

The implementation of a ranking principle for web-pages suitable for educational purposes is an important step for shifting the research for RS and IR methods in TEL to the Web. The design of such principle faces many problems. As well as common problems related to the scoring of web-pages for a query, an educational scoring is more complicated because of the many educational aspects that must be considered (Verbert et al., 2012). A good educational ranking principle should score higher those web-pages that present some educational aspects aligned with the user's educational needs. Instructors are responsible for the design and planning of the instruction, and they usually establish the educational requirements of



Estivill-Castro, V. and Marani, A.
Towards the Ranking of Web-pages for Educational Purposes.
DOI: 10.5220/0007586300470054
In *Proceedings of the 11th International Conference on Computer Supported Education (CSEDU 2019)*, pages 47-54
ISBN: 978-989-758-367-4

a course and materials. Hence, we review some studies on the teaching knowledge theory. This review suggests us what elements are relevant for scoring the teaching traits of a web-page.

Following a layered implementation of the principle, it is critical to scoring educational materials, in our case web-pages, according to instructional requirements first (Limongelli et al., 2013). Other educational elements can be introduced later for further improvement of the method. Once we identify those teaching requirements, we address some considerations on the structure of web-pages and possible structured information we can extract from them. Next, we present our proposal of a ranking principle, called *Educational Ranking Principle* (*ERP*) for scoring a web-page against the teaching context or requirements of a user. Finally, we conduct an offline experiment to validate our approach against traditional IR methods and Google. We build a dataset of web-pages labelled according to their suitability for teaching a concept in a particular educational context. Through an online survey, instructors defined a teaching context, searched for web-pages within the context and rated the top 10 web-pages that Google retrieved for a search. We report the tangible improvement of the quality of the ranking produced by our *ERP* than several baselines, including Google.

## 2 INFORMATION RETRIEVAL IN TECHNOLOGY ENHANCED LEARNING

The research areas of IR and RS have widely studied how to assist users in the retrieval of relevant goods and services from information sources. We can summarise that the problem of IR methods is to analyse the content of items in a dataset to retrieve relevant items to the user query. Instead, RS look at the users' rating and preferences of items to suggest new items without a user query and, generally, without full access to the content of the items. Hence, those systems use different approaches to support the discovery of items of interest.

The importance for the recommendation and retrieval of educational resources from the Web has significantly increased. Researchers propose recommender systems of Learning Objects for learners based on Collaborative Filtering (CF) and using ontologies. Maffon et al. (2013); Rodríguez et al. (2013); Limongelli et al. (2016b) are several approaches to handle and retrieve Learning Objects incorporating users' preferences, course structure or learning styles, where the focus is on a student user. We observe that the use of an ontology for the representation of the concept map of a course is a common element for expanding the user query or ranking resources.

However, some issues remain for the accurate recommendation of Learning Objects. Bozo et al. (2010) find that the Learning Object metadata fields are too complicated and ambiguous for use by a recommender system. Also, they discovered that the annotation of Learning Objects metadata is difficult to complete, and the metadata fails to represent pedagogical characteristics of the content. We concur with their conclusion about the lack of pedagogical descriptors in existing Learning Object metadata schema, but we need to address this problem differently. We can not expect to have web-pages with any educational metadata, so we can not rely on educational features of web-pages for scoring their usefulness.

We find that most of RS and IR proposals in TEL deal with items offering educational metadata (i.e. Learning Objects). Instead, our approach offers a fresh start for ranking web-pages for education. While we gain several insights on which Learning Object metadata can help in retrieval purposes, the scoring method must be innovative because we wish to score web-pages. When we look at surveys of IR and RS in TEL, we hardly find studies that consider web-pages (Drachsler et al., 2015). Even though the TEL research area is starting to look with interest at the Web (Estivill-Castro et al., 2018), the migration from Learning Object to web-pages is not easy (Estivill-Castro et al., 2018). Using Learning Object, we can assume that the resources we wish to present to the users are annotated, which is critical for scoring purposes. Unfortunately, web-pages do not generally offer any information of this kind. To the best of our knowledge, present solutions in TEL do not offer educational-based ranking of items with no educational metadata (Drachsler et al., 2015; Jensen, 2017). Since contextual information benefits web information retrieval, and educational studies have identified some important features of teaching contexts or resources in general, we propose here a methodology to combine these two research fields together to enable the recommendation and retrieval of web-pages for teaching.

For the definition of a teaching context, we look into the studies of the teaching knowledge theory (Voogt et al., 2013). These studies can provide us with some information on which educational aspects influence the selection and creation of learning resources. Our *ERP* aims to score the usefulness of a web-page according to such aspects. Following our analysis of the teaching knowledge (Voogt et al., 2013) and its applications (De Medio et al., 2016a; Limongelli et al., 2016c), *ERP* uses information about our understanding of a teaching context formed by the following 5 attributes: *Prerequisite Knowledge* (PK), *Concept Name* (CN), *Course Title* (CT), *Starting Knowledge* (SK) and *Target Knowledge* (TK). We expect that this information has potential for scoring the usefulness of a document for a teaching context, especially for documents with no educational metadata like web-pages. We already have found some applications of part of this information for comparing the performance of recommender systems in TEL (Lombardi and Marani, 2015), and building a new large dataset of educational resources from MOOCs. The result is called DAtaset of Joint Educational Entities (DAJEE) (Estivill-Castro et al., 2016). Exploiting such data, in particular the data in DAJEE, other studies have been able to address some relevant problems such as i) the automatic discovery of prerequisites of educational resources (De Medio et al., 2016a), ii) potential new recommendation methods for instructors (Limongelli et al., 2016a; De Medio et al., 2016b), iii) enrichment of the description of teaching resources Dessì et al. (2018); Limongelli et al. (2017). We show that this information holds the potential for implementing an educational ranking principle of web-pages for teaching.

# 3 SCORING WEB-PAGES FOR EDUCATION: *ERP*

*ERP* aims to rate web-pages according to some aspects of the teaching context. The rating should reflect the suitability of a web-page for teaching a particular concept in the specific instruction context. Our principle does not focus on the ranking of the pages according to a topic only; Google and other IR systems already do that with a remarkable performance. We want to challenge top IR methods from an educational perspective. The main problem we want to address is the sorting of a set of web-pages according to their suitability for teaching. A web-page, considered as an educational resource, is expected to explain a concept but, also, to refer to some fundamental knowledge around it (e.g. prerequisite knowledge) and be appropriate for the target students. and the detection of educational attributes by text-analysis only. Hence, the critical problem we address is the ranking of web-pages according to the educational context of the instructor, without any help from educational metadata about the resources.

## 3.1 Structured Information from Web-pages

Usually, for ranking web-pages, the body is the part of the page that is analysed unless some metadata is available (Pérez-Agüera et al., 2010). However, within the body of a web-page, we can find some tags that may contain a different kind of information. For example, the links (usually expressed by the HTML tag *a*) can have as text the name of other related concepts, while the headers should highlight important concepts of the web-page. For this reason, we want to distinguish the texts that come from the following four parts of a web-page: *title*, *body*, *links* and *highlights*. The results we present later in Section 4.2 indicate a superior accuracy performance of BM25F when compared to TFIDF. We believe that our *ERP* shall be a structured method like BM25F, but educationally oriented to achieve further progress and higher accuracy of the scoring phase. The problem is how to i) extract those four sections from web-pages, and ii) combine the information coming from these sections with the teaching context.

Table 1: HTML tags for a structured fetching of the content of web-pages. For each of the four parts of a web-page analysed by *ERP*, we indicate the HTML tags that we use for composing them.

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

We identify those four parts of a web-page by associating them to some HTML tags. Table 1 shows the four sections that the *ERP* can elaborate, and the corresponding HTML tags from where we extract the text populating the four sections. At this stage of the research, we only perform the removal of stop words and stemming of the texts using the Porter's stemming algorithm (Porter, 1980) before running any scoring method.

## 3.2 Matching the Attributes of the Teaching with Web-pages: The Expectancy Appearance Matrix (EAM)

Each attribute of the teaching context represents particular information about the educational requirements of a user. We defined four components of a web-page, and the attributes of the teaching context can have a different presence in each of those sections of a web-page. For the ranking process, *ERP* considers five attributes, and it analyses their appearance in the text fragments of the four sections of a web-page. To implement such a mechanism, we propose the *Expectancy Appearance Matrix* (EAM). We base this mechanism on the expectation that an attribute of the teaching context appears in a section of the web-page. The EAM reflects the covariance that an attribute is found in a section of the web-page. The rationale is that an attribute is more likely to appear in a part of a web-page instead of others. but we aim to reward web-pages which have elements of the teaching context in the right section, where those elements should be. A section of the web-page expresses a content with a specific meaning (for example, the section *links* mostly refers to related concepts, while the section *highlights* shall hold content about important concepts). A weighting method based on EAM should filter noise when an attribute of the teaching context repeatedly appears in a section of the web-page where we do not expect to find it. For example, we can consider noise a situation where a web-page presents a high frequency of the attribute *Concept Name* in the *links* section. We hypothesise that the web-page itself explains the concept without referring to other material to explain it. Following this example, we want to reward more a web-page which has a high frequency of the attribute *Concept Name* in the *title* and *body* sections. EAM assists *ERP* in better ranking web-pages by looking for each piece of the teaching context in the right section. Formally, EAM is a 4x5 matrix, where the rows are the four components of the web-pages analysed by *ERP*, and the columns are the attributes of the teaching context. The element $a_{ij} \in EAM$ is a weight expressing the expectancy that the *i*-th section of the web-page contains the *j*-th attribute of the teaching context.

### 3.3 Formulation of *ERP*

The purpose of *ERP* is scoring web-pages contrasting their content with the text of the attributes of the teaching context. *ERP* performs the matching of the texts by analysing the content of the four sections of a web-page. We base the scoring of each section on the TFIDF score, which is a potent method of scoring the relevance or similarity of a text for a query. In fact, even if the BM25F approach is a structured method, it still uses TFIDF-based scores for scoring a query for each section of the web-page. Roughly, BM25F changes the formulation of the term-frequency function to take into account the structure of the web-pages, while it maintains the IDF score. Similarly, we propose a TFIDF scoring of the four sections of a web-page in combination with a weighting system based on the EAM. As we introduce in Section 3.2, a column $j$ of EAM, called *Importance Vector* $IV_j$, indicates the expectancy that an attribute of the teaching context appears in the four sections of the web-page. The expectancy value $a_{ij}$ of the EAM weights the TFIDF scores of the terms of the $j$-th attribute in the $i$-th section of the web-page. For the $j$-th attribute, we formally define $IV_j$ as follows.

$$IV_j = \left\langle a_{title,j}, a_{body,j}, a_{links,j}, a_{highlights,j} \right\rangle,$$

where $a_{ij}$ is an element of EAM and $j \in \{\mathrm{CN}, \mathrm{CT}, \mathrm{PK}, \mathrm{SK}, \mathrm{TK}\}$. We normalise the frequency values with the number of words of the text. In practice, we apply the traditional TFIDF formula with normalised frequencies:

$$TFIDF(attText, secText) = \sum_{term \in attText} \frac{freq(term, secText)}{length(secText)} \cdot IDF(term)^2 \quad (1)$$

where $length(secText)$ returns the number of words in the text, in practice, the sum of the frequencies of all the terms in *secText*. The $IDF(term)$ function is defined as follows:

$$IDF(term) = 1 + \log \frac{total\ number\ of\ documents}{docFreq(term) + 1}. \quad (2)$$

We finally use this TFIDF function for building a *TFIDF Vector TFIDF-V* of the terms of the $j$-th attribute in the four parts of the web-page, namely:

$$\begin{aligned} TFIDF\text{-}V_j = \\ &\langle \mathrm{TFIDF}(attText_j, titleText), \\ &\mathrm{TFIDF}(attText_j, bodyText), \\ &\mathrm{TFIDF}(attText_j, linksText), \\ &\mathrm{TFIDF}(attText_j, highlightsText) \rangle. \end{aligned}$$

Finally, for each attribute of the teaching context, *ERP* computes the sum of the dot products of the TFIDF vectors *TFIDF-V* with the importance vectors *IV*. We normalise the score with the sum of the dot products of the importance vectors with the vector of maximum IDF values *IDF-V*:

$$ERP = \frac{\sum_{j \in \{\mathrm{PK,CN,CT,SK,TK}\}} IV_j \cdot TFIDF\text{-}V_j}{\sum_{j \in \{\mathrm{PK,CN,CT,SK,TK}\}} IV_j \cdot IDF\text{-}V} \quad (3)$$

Given a collection of documents, the vector *IDF-V* consists of four entries (one for each section) which are the highest IDF value squared, $IDF_{max}{}^2$, because the TFIDF function has its IDF component squared. Ideally, the highest IDF happens when there are no body-texts in the dataset that contain a term. Following Formula 2, the value of $IDF_{max}$ is $1+\log(numDocs)$.

By construction, the co-domain of *ERP* is the interval $[0,1]$. The TFIDF scores are positive and bounded above by $IDF^2_{max}$, since the term-frequency values are normalised to 1. At most, the numerator is the sum of dot products between *IDF-V* and the importance vectors. The denominator is such sum; so, dividing numerator and denominator the result is 1 that is the maximum value of *ERP*. In the case no sections of a web-page contains any term of the teaching context attributes, the TFIDF vectors are vectors of zeros because the term-frequency values are zero. In this case, the sum of the dot products at the numerator returns zero, while the denominator is still greater than 0 (the importance vectors are positive vectors with at least one element greater than zero), so *ERP* is equal to 0.

This formulation of the *ERP* does not include all of the attributes of the teaching context (i.e. *Difficulty Level* and *Education Level*). The problem of introducing them in the current *ERP* is the definition of proper values for their importance vectors in the EAM.

# 4 EVALUATION OF *ERP*

We performed a data collection phase for building a dataset of web-pages rated in teaching contexts. The ratings reflect the suitability of a web-page for teaching a concept in a context defined by instructors themselves. Following good practices for scoring the usefulness of web-pages(Mao et al., 2016), we implemented an online survey where instructors can label the usefulness of web-pages. We ask instructors to i) define a teaching context of their interest, which includes a concept map, ii) formulate a query for retrieving web-pages for a concept of the concept map, and iii) rate the usefulness of the retrieved web-pages for teaching the concept in the defined teaching context. Our online survey interrogates Google[1] and presents the first ten items in a random order for avoiding any bias due to the system's presentation order. After an automatic quality control, the dataset hosts a total of 614 web-pages rated by instructors who conducted 66 web searches about 23 teaching contexts.

The goal of this evaluation is to prove that *ERP* scores web-pages more accurately than current practice and baseline methods. We use position-based and prediction-accuracy measures (Shani and Gunawardana, 2011) for proving the higher accuracy of our method than TFIDF, BM25F and Google. The measures we use in this experiment are Mean Average Precision (MAP) and precision scores at the top 1 (P@1), top 3 (P@3) and top 5 (P@5) positions (Shani and Gunawardana, 2011).

We base the statistical significance of our analysis on paired $t$-tests. Given the size of our sample data, we can apply paired $t$-tests for the analysis of the stronger performance of our method compared to each baseline. We run one paired $t$-test for each measure and baseline method. The null hypothesis $H_0$ is that the performance of the baseline is higher than our *ERP* on average. The software R (R Core Team, 2016) runs the $t$-tests.

## 4.1 Values for the EAM

Similarly to other IR scoring methods, the discovery of the optimal setting of the EAM is an interesting challenge that usually requires an extensive study and application of machine-learning machinery over a substantial number of queries Pérez-Agüera et al. (2010). Since *ERP* is new proposal and our dataset is small for undertaking such task Pérez-Agüera et al. (2010), we devised an ad-hoc algorithm to identify suboptimal values for *ERP*. As result of the algorithm, we propose the following values for the EAM rounded to one decimal:

Table 2: The EAM for running *ERP*.

| | **CN** | **CT** | **PK** | **SK** | **TK** |
|---|---|---|---|---|---|
| **Title** | 0.9 | 0.2 | 0.9 | 1.0 | 0.1 |
| **Body** | 1.0 | 0.8 | 0.7 | 0.1 | 0.2 |
| **Links** | 0.0 | 0.2 | 0.0 | 0.1 | 0.4 |
| **Highlights** | 0.2 | 0.0 | 0.1 | 0.2 | 0.4 |

Table 3: Results of the performance of TC-informed TFIDF, TC-informed BM25F, Google and *ERP* for ranking the results of 66 web searches in our dataset. This table reports the accuracy performance of the methods according to MAP and the average values of P@1, P@3 and P@5. In bold, we highlight the highest score recorded in the experiment.

| **Metric** | **TC-informed TFIDF** | **TC-informed BM25F** | **Google** | ***ERP*** |
|---|---|---|---|---|
| MAP | 0.579 | 0.606 | 0.627 | **0.683** |
| P@1 | 0.485 | 0.545 | 0.545 | **0.712** |
| P@3 | 0.5 | 0.525 | 0.561 | **0.601** |
| P@5 | 0.485 | 0.512 | 0.527 | **0.567** |

We are aware that such solution is a local optimal solution and not the optimal solution for *ERP*. However, this is sufficient for our experiment. If we

[1]Google is queried by using the Google Custom Search service expanded to the entire web.

Table 4: Results of the paired $t$-tests of the AP measure on 66 queries. We compare *ERP* with the baseline methods and systems. The *Mean of the differences* column is the average difference of the MAP values.

| ***ERP* VS** | ***t*-value** | **95% Confidence interval of mean differences** | **Mean of the differences** | ***p*-value** |
|---|---|---|---|---|
| TF-IDF (TC-informed) | 2.846 | 0.032 | 0.077 (+12.71%) | 0.003 |
| BM25F (TC-informed) | 2.383 | 0.017 | 0.056 (+8.93%) | 0.010 |
| Google | 4.521 | 0.066 | 0.104 (+17.96%) | 1.335E-05 |

Table 5: Results of the paired $t$-tests of the P@1 measure on 66 queries. We compare *ERP* with the baseline methods and systems.

| ***ERP* VS** | ***t*-value** | **95% Confidence interval of mean differences** | **Mean of the differences** | ***p*-value** |
|---|---|---|---|---|
| TF-IDF (TC-informed) | 2.372 | 0.049 | 0.17 (+30.64%) | 0.01 |
| BM25F (TC-informed) | 2.494 | 0.055 | 0.17 (+30.64%) | 0.008 |
| Google | 3.363 | 0.115 | 0.227 (+46.8%) | 0.0006 |

discover that *ERP* performs better than the baselines with this non-optimal EAM, the validation is sufficient. Then, the discovery of the optimal EAM is a problem that is worth to be further investigated in future studies when a larger dataset is available.

## 4.2 Results and Discussions

We now present the results of the performance of *ERP* against Google, BM25F and TFIDF using the data of 66 web searches.

We find that our *ERP* performs better than all the baselines. Also, the performance measures of *ERP* are remarkably higher than the benchmarks. The MAP, P@1, P@3 and P@5 measures indicate a more tangible difference, especially MAP and P@1 which are extremely relevant for the evaluation of IR methods (Shani and Gunawardana, 2011). Table 3 reports the values of the performance measures for all the baselines and *ERP*. Following the measure P@1, the best baselines are TFIDF and BM25F with a P@1 value of 0.545 while *ERP* achieves a remarkable 0.712. *ERP* increases the performance of 0.167 which is 30.64% higher than the performance of the best baselines. Looking at the MAP measure, we find that BM25F is the most accurate baseline with a MAP value of 0.627 compared to 0.683 of *ERP*. In this case, the difference between the MAP values is 0.056, and *ERP* achieves a MAP score that is 8.93% higher than BM25F. We further analyse the results with a set of paired $t$-tests, one test for each measure and baseline, for exploring the statistical significance of the results. We need a minimum $t$-value of 2.000 for rejecting the null hypothesis at .05 significance level for the degree of freedom equal to 65.

Table 4 reports the paired $t$-tests for the comparison of the AP measure of *ERP* against TFIDF, BM25F and Google. The column *Mean of the differences* is, in essence, the difference of the MAP values of *ERP* and the baseline approaches. Our *ERP* remarkably increases the MAP performance. The low $p$-values (under 0.05) and the enough robust $t$-values confirm the rejection of the null hypothesis at 0.05 significance level for all the baselines. Therefore, we can say that our *ERP* ranks useful web-pages at higher positions than the three baselines.

We now look at the P@1 performance of the methods reading Table 5. This measure is rigorous since it is either 1 or 0; it says if the web-page at the first position of the ranking is useful or not. On average, the experiment shows that *ERP* performs better than all the baselines as the $p$-values under .05 support. Also, the paired $t$-tests reject the null hypothesis for all the baselines, indicating the strong performance of *ERP*. On average, the estimate for the differences of P@1 between *ERP* and the baselines is at least 0.05 with 95% confidence.

In the case of P@3, Table 6 shows that *ERP* does not perform well against BM25F, while the null hypothesis is rejected for the other baselines. The lower-bound of the confidence interval of the mean differences is marginally in favour of BM25F. Also, the $p$-value is not low enough for confirming the average of the differences of 0.045 that we find in our experiment. In this case, we compute the paired $t$-test that the mean of the P@3 measure of BM25F is higher than *ERP*. The results are not positive, with a negative $t$-value (-1.209) and a high $p$-value (0.885). This result confirms that *ERP* is more likely to perform better than BM25F, but the current experimental data are not strong enough for a statistical confirmation. We can, however, generalise the very positive results we obtain from *ERP* against TFIDF and Google.

The last analysis is about the P@5 measure, and

Table 6: Results of the paired $t$-tests of the P@3 measure on 66 queries. We compare *ERP* with the baseline methods and systems.

| ***ERP* VS** | ***t*-value** | **95% Confidence interval of mean differences** | **Mean of the differences** | ***p*-value** |
|---|---|---|---|---|
| TF-IDF (TC-informed) | 2.156 | 0.017 | 0.076 (+14.48%) | 0.017 |
| BM25F (TC-informed) | 1.209 | -0.015 | 0.04 (+7.13%) | 0.116 |
| Google | 2.928 | 0.043 | 0.10 (+20.20%) | 0.002 |

Table 7: Results of the paired $t$-tests of the P@5 measure on 66 queries. We compare *ERP* with the baseline methods and systems.

| ***ERP* VS** | ***t*-value** | **95% Confidence interval of mean differences** | **Mean of the differences** | ***p*-value** |
|---|---|---|---|---|
| TF-IDF (TC-informed) | 2.149 | 0.012 | 0.055 (+10.74%) | 0.018 |
| BM25F (TC-informed) | 1.688 | 0.0004 | 0.039 (+7.59%) | 0.048 |
| Google | 3.791 | 0.046 | 0.082 (+16.91%) | 0.0002 |

Table 7 reports the outcome of the paired $t$-tests for this measure. Only against BM25F we cannot reject the null hypothesis. When comparing *ERP* to BM25F, we have a $t$-value of 1.688 which is not sufficient to establish statistical significance. However, the $p$-value is slightly lower than 0.05. Hence, we can expect that, on average, *ERP* has a better P@5 accuracy than BM25F like we recorded in this experiment.

## 5 CONCLUSIONS AND FUTURE WORK

We investigated the problem of an educational oriented ranking of web-pages for teaching. We propose a new ranking principle, called *ERP*, which can rank web-pages without any educational information of their content. Our work opens a new direction for IR in TEL to the web instead of Learning Object Repositories only. Out of three baseline methods and four accuracy measures, our *ERP* is a more reliable scoring method than current practice. Our experiment is based on web-pages retrieved by Google search engine, and instructors defined the teaching contexts according to their real experience. Hence, *ERP* can already practically assist instructors to detect those web-pages that are better placed for teaching in their contexts. Moreover, it is important to highlight the remarkable increase of MAP and P@1 with *ERP*. These results are positive and statistically significant as per the reported paired $t$-tests. *ERP* shows a MAP value higher than BM25F, the best performing baseline, of 8.93%, while the improvement for P@1 is of 30.64%. These results are impressive and strongly support our proposal.

The experiment of this study is limited to scoring 10 web-pages at the time. We need to investigate its performance and accuracy when scoring an extremely large number of web-pages, a scenario that better reflects the size of the web. In this sense, some pilot studies report the meaningful enrichment of TEL datasets with semantic information Limongelli et al. (2017). Such semantic information may assist our *ERP* when scoring an extremely large amount of web-pages at the same time.

## REFERENCES

Bozo, J., Alarcón, R., and Iribarra, S. (2010). Recommending learning objects according to a teachers' contex model. In *Sustaining TEL: From innovation to learning and practice*, pages 470–475. Springer.

Brin, S. and Page, L. (2012). Reprint of: The anatomy of a large-scale hypertextual web search engine. *Computer networks*, 56(18):3825–3833.

De Medio, C., Gasparetti, F., Limongelli, C., Lombardi, M., Marani, A., Sciarrone, F., and Temperini, M. (2016a). Discovering prerequisite relationships among learning objects: a coursera-driven approach. *Int. Conf. on Web-Based Learning*, pages 261–265. Springer.

De Medio, C., Gasparetti, F., Limongelli, C., Lombardi, M., Marani, A., Sciarrone, F., and Temperini, M. (2016b). Towards a characterization of educational material: an analysis of coursera resources. *Int. Symposium on Emerging Technologies for Education*, pages 547–557. Springer, Cham.

Dessì, D., Fenu, G., Marras, M., and Recupero, D. R. (2018). Coco: Semantic-enriched collection of online courses at scale with experimental use cases. In *World Conference on Information Systems and Technologies*, pages 1386–1396. Springer.

Drachsler, H., Verbert, K., Santos, O. C., and Manouselis, N. (2015). Panorama of recommender systems to support learning. In *Recommender systems handbook*, pages 421–451. Springer.

Estivill-Castro, V., Limongelli, C., Lombardi, M., and Marani, A. (2016). Dajee: A dataset of joint educational entities for information retrieval in technology enhanced learning. *39th Int. SIGIR Conf. on Research and Development in Information Retrieval*, pages 681–684. ACM.

Estivill-Castro, V., Lombardi, M., and Marani, A. (2018). Improving binary classification of web pages using an ensemble of feature selection algorithms. *Australasian Computer Science Week Multiconference*, page 17. ACM.

Jensen, J. (2017). A systematic literature review of the use of semantic web technologies in formal education. *British J. of Educational Technology*.

Kay, J., Reimann, P., Diebold, E., and Kummerfeld, B. (2013). Moocs: So many learners, so much potential... *IEEE Intelligent Systems*, (3):70–77.

Limongelli, C., Lombardi, M., and Marani, A. (2016a). Towards the recommendation of resources in coursera. *Intelligent Tutoring Systems: 13th Int. Conf., ITS, Zagreb, Croatia, June 7-10*. volume 9684, page 461. Springer.

Limongelli, C., Lombardi, M., Marani, A., and Sciarrone, F. (2013). A teaching-style based social network for didactic building and sharing. In *Artificial Intelligence in Education*, pages 774–777. Springer.

Limongelli, C., Lombardi, M., Marani, A., Sciarrone, F., and Temperini, M. (2016b). A recommendation module to help teachers build courses through the moodle learning management system. *New Review of Hypermedia and Multimedia*, 22(1-2):58–82.

Limongelli, C., Lombardi, M., Marani, A., Sciarrone, F., and Temperini, M. (2016c). Concept maps similarity measures for educational applications. *Intelligent Tutoring Systems: 13th Int. Conf., ITS, Zagreb, Croatia, June 7-10*, pages 361-367. Springer.

Limongelli, C., Lombardi, M., Marani, A., and Taibi, D. (2017). Enrichment of the dataset of joint educational entities with the web of data. *Advanced Learning Technologies (ICALT), 2017 IEEE 17th Int. Conf.* , pages 528–529. IEEE.

Lombardi, M. and Marani, A. (2015). A comparative framework to evaluate recommender systems in technology enhanced learning: a case study. In *Advances in Artificial Intelligence and Its Applications*, pages 155–170. Springer.

Maffon, H. P., Melo, J. S., Morais, T. A., Klavdianos, P. B., Brasil, L. M., Amaral, T. L., and Curilem, G. M. (2013). Architecture of an intelligent tutoring system applied to the breast cancer based on ontology, artificial neural networks and expert systems. *ACHI, Sixth Int. Conf. on Advances in Computer-Human Interactions*, pages 210–214.

Maloney, S., Moss, A., Keating, J., Kotsanas, G., and Morgan, P. (2013). Sharing teaching and learning resources: perceptions of a university's faculty members. *Medical education*, 47(8):811–819.

Mao, J., Liu, Y., Zhou, K., Nie, J.-Y., Song, J., Zhang, M., Ma, S., Sun, J., and Luo, H. (2016). When does relevance mean usefulness and user satisfaction in web search? *39th Int. SIGIR Conf. on Research and Development in Information Retrieval*, pages 463–472, ACM.

Pérez-Agüera, J. R., Arroyo, J., Greenberg, J., Iglesias, J. P., and Fresno, V. (2010). Using bm25f for semantic search. *3rd Int. semantic search workshop*, page 2. ACM.

Porter, M. F. (1980). An algorithm for suffix stripping. *Program*, 14(3):130–137.

R Core Team (2016). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rodríguez, P. A., Tabares, V., Mendez, N. D. D., Carranza, D. A. O., and Vicari, R. M. (2013). Broa: An agent-based model to recommend relevant learning objects from repository federations adapted to learner profile. *IJIMAI*, 2(1):6–11.

Shani, G. and Gunawardana, A. (2011). Evaluating recommendation systems. In *Recommender systems handbook*, pages 257–297. Springer.

Verbert, K., Manouselis, N., Ochoa, X., Wolpers, M., Drachsler, H., Bosnic, I., and Duval, E. (2012). Context-aware recommender systems for learning: a survey and future challenges. *Learning Technologies, IEEE Transactions on*, 5(4):318–335.

Voogt, J., Fisser, P., Pareja Roblin, N., Tondeur, J., and van Braak, J. (2013). Technological pedagogical content knowledge–a review of the literature. *J. of Computer Assisted Learning*, 29(2):109–121.