BiblioMiner: Answer to exam questions based on the indexing of bibliographic elements

Carlos André Cunha Neves Abrantes
carlos.abrantes@ist.utl.pt

Instituto Superior Técnico, Lisboa, Portugal
October 2015

Abstract

Every year, students learn the documents in the bibliography of a discipline and are tested in a final exam which aims at the evaluation of their acquired knowledge. For students it is necessary to have quick access to the most relevant texts related to a question they do not know the answer to. This project extracts indicators of the bibliography such as the words of the summary, toc, TF and IDF, and builds a dataset based on an online survey of questions and answers. This dataset is used to train and test a “Learning to Rank” algorithm that uses linear, logistic and SVM models. The results obtained show that the logistic model, with kernel Epachnenikov and shuffled sampling, is the best model, due to its ROC curve proximity with the best model point, maintaining good levels of precision and F1 when compared with the remaining combinations of models, kernels and shuffling types.

Keywords: question-answering, passage retrieval, L2R, structure of books, keyword selection

1. Introduction

There are about 139,000 educational institutions [1] in U.S.A.. Richard Arum et al. [2] found that on average each American college student spends about 24 hours a week in school activities including 9 hours studying alone. BiblioMiner, will be an search engine tool that will look for relevant documents of the bibliography according with a specific question asked by the student. This work aims to implement and evaluate the BiblioMiner system that aims to: monetize user study time; be transversal to any knowledge domain and ensure that the returned texts of bibliography are important to answer the question.

The BiblioMiner is a system composed of the following steps: i) create a dataset adapted to a discipline using an online survey of questions and answers filled by an expert in the discipline; ii) extract the indicators of the dataset; iii) apply Learning to Rank algorithms to figure out which of these indicators are more relevant; iv) use the algorithm with better performance to provide relevant answers to user questions asked. This dissertation is divided into seven chapters, in the first chapter is described the framework of the dissertation, the objectives and the methodology used. In chapter 2
are covered basic concepts essential to the interpretation of the related work and the proposed solution. Chapter 3 consists in the description of QR systems, indexing, meta-data extraction and selection of keywords that were helpful to understand the state of the current art. The chapter 4 corresponds to the proposal which serves as a model to the prototype deployed. In Chapter 5 are described some of the outlines of the deployment options made. Chapter 6 presents the evaluation of the results and finally the chapter 7 contains the conclusions and future work.

2. Basic Concepts

2.1. Passage Ranking systems

The Passage Retrieval systems use queries to search relevant passages in collections of documents [6, 12, 13]. This systems are composed in two parts: passaging and passage ranking. Passaging module makes the segmentation, filtering and database storing. The module of Passage Ranking defines the set of candidate passages based on criterias and having paying attention to the relevant indicators.

The purpose of these systems is to discover the passages from the collection of documents that may be more important to answer a question. The passaging module is well established by authors such as Hussain et al. [3], Liu et al. [5] and Veravalli et al. [9] who substitute separation marks such as period, blank lines or fixed windows of M bytes to segment the passages by white spaces. The passages are then indexed and stored in a repository such as a database.

The ranking passage module has the intention of sorting the passages according to its punctuation score. Each passage is associated with a score assigned by a scoring function (FP) which is constituted by a set of relevant indicators.

A few common indicators are: CS, BM25, PageRank, cosine similarity, TF, IDF and TF*IDF, that are present in the related work section.

2.2. Learning to Rank

In order to determine the weights assigned to each indicator of the previous chapter, can be applied techniques L2R [8]. L2R is an application of machine learning (ML) to document retrieval problems. These systems detect patterns that indicate if a response is better than another. In this context, the aim of such systems is to find automatically, a scoring model using a collection of data. Each data element of the set (dataset) is composed by a question \( q_i \), an array of indicators \( x \), that represents a document, and a ground of truth \( y_i \) associated with that vector. The indicator array is obtained from a function that aims to extract relevant properties between the question and the document. The ground of truth is restricted to the range of values that rates the relevance of the document.

In Figure 1 is shown an L2R system. The dataset is divided in such a way that it can be used by the training data, test data and validation data. The learning module task aims to find a model that suits the solution of the problem, based on the training data. The test data consists of the dataset elements without the ground of truth. To determine whether the model generated by the learning module correctly maps a label with an array of indicators, the model is tested with test data leading to a prediction.

Is possible to evaluate the performance of an event, using a loss of function (LF - loss function) that makes the error measurement between the ground of truth of the dataset and the predictive model calculated by \( h(x) \). The validation data is used by the training algorithm to adjust its hyper
parameters such as the number of algorithm iterations. It is through these parameters that the L2R systems learn the best way to combine the indicators array in order to minimize the error.

To make the evaluation of the generated L2R system can be used the experimental methodology of cross-validation. This method assumes that there is a set of N data elements which can be divided into blocks B constituted by k elements. Can be used N-2k elements to training (training data), k elements to test (test data) and k elements to validate the model (data validation). This process is repeated B times in a rotative way varying the B blocks by the different sets. The comparison between systems may be made by making a performance evaluation of the system in the end of each iteration of the cross validation methodologies.

3. Related Work

3.1. Learning to Rank

Verberne et al. [10, 11] have created a why QA-system which utilizes 37 relevance indicators. Some of the most relevant indicators were the title of the document and the synonyms of the terms of the question (it turned out that these synonyms are usually contained in the passage with greater score). Since the goal of Verberne et al. is to compare the behavior of various systems L2R maintaining the same training data, tests using the this 3 categories have been made: point-wise, pairwise and listwise. These categories utilizes the ML algorithms: i) Naive Bayes, Support Vector Classification (SVC), Support Vector Regression (SVR), logistic regression ii) Ranking SVM iii) genetic algorithm. The why-questions used in the training data were extracted from Wikipedia, Webclopedia¹ e Answers.com.

The systems that show the best behave are the ones with machine learning algorithm SVR and the genetic algorithm. The authors [11], authors point out that the method of Logistic Regression achieve good results using the default values. However, the algorithms that performed better than the previous stated that its parameters were adjusted manually.

4. BiblioMiner

The BiblioMiner system combines techniques used in the past by Passage Retrieval, Keyword Selection, Indexing and meta-data extraction using the structure of books and L2R systems.

The architecture of the system represented in Figure 2 uses an L2R algorithm to score all the passages from books and slides the literature of a discipline. The documents of the bibliography are indexed by the module of Passage Retrieval (PR). For each exam question, all candidate passages elected by the PR module are collected.

In an online poll, users associate with each pair question-passing a degree of ground of truth. After voting, the results are preprocessed. In the extraction module the arrays of indicators are extracted and associated with the respective ground of truth. This data is going to be a part of the dataset used for the cross-validation module. The

Figure 1: System Learning to Rank, adapted from Liu et al. [4]

Figure 2: System architecture with the L2R algorithm.
iterations of the L2R system over these data results in a learning model. The learning module with enhanced performance (Performance modules L2R and MAX) is used by the system to return the most relevant passages concerning a question of the user. In the next sub-sections, the creation of the dataset and modules L2R system will be emphasized.

4.1. Dataset creation

L2R systems need training data in order to create a hypothesis in the learning module. In the case of this project, it was not found on previous works of the state of the art, questions associated to academic slides, summary and book. In order to accomplish our study it was necessary to develop an online vote system that assists on the creation of the dataset. The creation of the dataset will be made in three steps: classification of the relevant passages (represented by the PR modules and online voting in Figure 2); preprocessing the content of the online votation (represented by the arrow between the module of online voting and extraction and associated indicators to the vote in Figure 2); extraction of the online voting indicators and association to the respective ground of truth (represented by the extraction module and associated indicators to the votation in Figure 2).

4.2. Passage relevance classification

To classify the relevance of the passages regarding an examination question, we can use an online voting system that allows to classify the passages as relevant / irrelevant. The vote was restricted to a team of evaluators. The bibliography of the elements used are from SHELF, a repository of books and slides of the subject Management for Telecommunications and Information Technology. The summaries have been stored by hand.

The Passage Retrieval module (PR) in Figure 2 All books and slides will be indexed using the Whoosh system, a text indexer. A list of questions from the classes’ previous exams will be manually created. For each question inserted in the Whoosh system, K passages will be returned (utilizing the TF-IDF return function), which will then be on display online, so system evaluators can vote them as relevant/irrelevant. At the end of this process, we will have a question associated to a set of relevant passages and to a set of irrelevant ones.

4.3. Online poll content pre-processing

The contents of the bibliography’s documents present words that are irrelevant to a research system. Several pre-processing techniques will be tested, such as the removal of stopwords2, stemming, normalization, syntactic parsing, passage size definition and number of words on each *.

4.4. Extraction and association of indicators to online poll

This project’s learning module will be trained to identify which elements of the vector of question-passage indicators are most important to recognize passages that answer correctly to a question. Each indicator in the vector expresses a property’s occurrence frequency. The indicator’s vector in this project has 21 indicators: Id1-term in bold in the passage ;Id2- term in italics in the passage ;Id3- term in summary ;Id4-

---

2Stopwords are a list of words without meaning in the context of a research.
exam term ;Id5- term in the index ;Id6- term in the index contents ;Id7- Minimum font size of the words in CS;Id8- Maximum size letter of the words in CS;Id9- Size medium font of the words in CS;Id10- TF sum of terms in CS ;Id11- IDF sum of terms in CS ;Id12- TF * IDF sum of the terms in CS ;Id13- Terms of CS that first appears on the slide ;Id14- Terms of CS appearing for the last time on the slide ;Id15- Set size selection;Id16- Cosine similarity between the question and the passage;Id17- Minimum distance of the terms of the CS slide;Id18- Maximum distance of the terms of the CS slide;Id19- Average distance of the terms of the CS slide;Id20- Area of these figures in the document and Id21- Number of pictures on the page.

After the closing of the polls of the online system, the textual elements are pre-processed and the question-passage indicators and their respective degree of truth are extracted.

4.5. Learning to Rank

The created dataset is utilized to train the L2R system’s learning algorithm. The following tasks will be executed to find the algorithm that presents the best possible results:
- Indicator set modification, or, testing combinations of indicators until it figures out which ones optimize the results.
- Alteration of the ML algorithm and its parameters.

The Large data rationale criteria assumes the indicator vector x contains enough information to predict y. Because of this, in this project, we can apply the multivariable linear regression and logistics models in the learning module since 21 indicators are utilized which come from the state of the art and consultation methods used in books.

5. Experimental validation

In this chapter is made the assessing of the performance of the proposed system in previous chapter, the followed methodologies and results obtained during the experimental validation. We used the dataset of discipline of Management 2008-2009 of the bachelor of Telecommunications and Information Technology of the Instituto Superior Técnico.

The comparative analysis between models is made as follows: for each model is chosen the combination of kernel, sampling and indicators that allow the system to obtain greater precision; the chosen models are compared in terms of precision and F1 in function of the number of elements of the training set; the chosen models are compared by overlapping the ROC curves and AUC.

5.1. Description of dataset of data

Was created an dataset based on bibliographic elements of the 2008/2009 year of the Management discipline of the course of Telecommunications and Information Technology. The slides consist of 409 pages and the exam questions consist of 602 questions drawn manually. To reduce the noise it was decided to remove all non-alphanumeric characters except white space as done by Lyu et al. [7]. The 3 evaluators take about 18.3 hours (1100 minutes) to classify the 602 questions of the online poll.

5.2. Chronology of the dataset

This project focuses on improving students productivity levels taking into account the quality and study time inherent to the learning the didactic content. During the online voting was monitored the overall time that the evaluators took to answer questions.

The evaluators take about 1100 minutes to review the entire dataset management. in Table 1, is found the way it was constructed the dataset over
Table 1: Relevant and non-relevant elements linked to the length of the online voting

<table>
<thead>
<tr>
<th>Average votation (pool duration minutes)</th>
<th>Relevant Elements</th>
<th>Non-Relevant Elements</th>
</tr>
</thead>
<tbody>
<tr>
<td>13</td>
<td>10</td>
<td>19</td>
</tr>
<tr>
<td>23</td>
<td>20</td>
<td>30</td>
</tr>
<tr>
<td>38</td>
<td>30</td>
<td>54</td>
</tr>
<tr>
<td>50</td>
<td>40</td>
<td>70</td>
</tr>
<tr>
<td>69</td>
<td>50</td>
<td>102</td>
</tr>
<tr>
<td>97</td>
<td>60</td>
<td>152</td>
</tr>
<tr>
<td>129</td>
<td>80</td>
<td>202</td>
</tr>
<tr>
<td>163</td>
<td>100</td>
<td>257</td>
</tr>
<tr>
<td>206</td>
<td>200</td>
<td>250</td>
</tr>
<tr>
<td>538</td>
<td>300</td>
<td>877</td>
</tr>
<tr>
<td>745</td>
<td>400</td>
<td>1231</td>
</tr>
<tr>
<td>1100</td>
<td>422</td>
<td>1986</td>
</tr>
</tbody>
</table>

The resultant system of this project need to have high level of precision (rate of returned elements which are relevant). The experimental validation was started by choosing the models with higher precision. To achieve this goal, it was necessary to make the necessary variations of kernel, sampling and indicators using feature elimination algorithms. Then we studied the F1 and precision behavior using a variable-length dataset's (Charts 3, 4, 6, 7) analyse the overlapping ROC curves for each model (Charts5 and 8) comparing the AUC values.

5.3. Evaluation of linear models, logistic and SVM

Throughout this chapter will be evaluated the ability of modeling the solution of the proposal problem using linear, logistic and SVM models. It will be considered the performance of the models using the average precision e F1 using 10-fold cross-validation L2R system.

To test the precision and F1 of L2R systems it was necessary to control the number of relevant dataset objects and not relevant, otherwise the results would be biased. Was tested the system behavior in two situations: Situation 1: Dataset R = NR (number of relevant elements = number of non-relevant elements) and Scenario 2: Dataset 2 R = NR (number of relevant elements is half the number of non-relevant elements).

The resultant system of this project need to have high level of precision (rate of returned elements which are relevant). The experimental validation was started by choosing the models with higher precision. To achieve this goal, it was necessary to make the necessary variations of kernel, sampling and indicators using feature elimination algorithms. Then we studied the F1 and precision behavior using a variable-length dataset's (Charts 3, 4, 6, 7) analyse the overlapping ROC curves for each model (Charts5 and 8) comparing the AUC values.

5.4. Situation 1: Dataset R = NR

5.4.1 Model selection, kernel and sampling data

The first model to be analyzed is the linear model, which considers the three types of sampling: linear, shuffled stratified. The kernels used in the linear model are: None, M5 prime, Greedy and T-test. Then, considering the same data set (R = NR) it was applied the logistic and SVM models. In this case the sampling is the same as the previous and the kernels are Dot, Radial, polynomial, Neural, Anova, Epachnenikov, Gaussian combination multiquadric. The results summary is present in the Table 2.

At this stage, we can conclude that with no further type of operation SVM model with kernel Dot and sampling Stratified would have been the model with higher precision using a dataset with balanced classes.

The it was use feature selection algorithms to obtain better results. Were applied backward elimina-
tion and forward validation. Removing some feature indicators it was possible to increase the precision of the linear and logistic models for the values 67.14% and 69.01%, respectively. These predictions were achieved after removed the id4 indicators, id10, and id16 the linear model and the indicators id15, id5 and id8 of the logistic model. The SVM model has not recorded any improvement using the feature selection algorithms.

5.4.2 Precision analysis by varying the size of the dataset

On this section, the main objective is to determine how the size of the training influences the precision and F1 in the situation R = RN. The graphical analysis of the different models allows the relative comparison between models, referring to Figure 3.

By evaluating the overall precision of evolution of different models, it can be concluded that the linear model has the highest precision in the range 20 to 80 driving elements. However, this model has inconstant precision curves over the range. The logistic model in comparison with other models, has the highest precision to the range 80 to 800 training elements. The SVM model does not stand out as to the precision of any of the previous models, but it is the one with greater precision homogeneity over the range 80 to 800 training elements.

Figure 3: Precision depending on the number of training elements of the set R = NR

5.4.3 ROC curve analysis

This section is intended to compare the ROC curves of the chosen models R = NR. This analysis allows you to determine which model best classifies the relevant slides as relevant and non-relevant sides as not relevant. As can be seen in Figure 3 models feature a similar curvature and above the line (0,0) a (1,1), this means that all models outperform the random model. However, the linear model is the one closer to the random line. In this case, the best classifier is not easily identifiable graphic perspective. Therefore find the best classifier using the AUC models. The order from best
Table 3: Elected models 2R = NR

<table>
<thead>
<tr>
<th>Model</th>
<th>Logistic</th>
<th>SVM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Linear</td>
<td>65.15%</td>
<td>77.78%</td>
</tr>
<tr>
<td></td>
<td>86.67%</td>
<td></td>
</tr>
</tbody>
</table>

The first model to be analyzed is the linear model, which considers the three types of sampling: linear, shuffled stratified. The kernels used in the linear model are: None, M5 prime, Greedy and T-test. Then, considering the same data set (2R = NR) it was applied the logistic and SVM models. In this case the sampling is the same as the previous and the kernels are Dot, Radial, polynomial, Neural, Anova, Epachnenikov, Gaussian combination multiquadric. The results summary is present in the Table 3.

At this stage, we can conclude that with no further type of operation logistic model with kernel Epachnenikov and sampling shuffled would have been the model with higher precision using a dataset with balanced classes.

5.5.1 Precision of analysis by varying the size of the dataset

On this section, the main objective is to determine how the training set size influence the precision and F1 of the different models chosen 2R = RN. Referring to Figure 6 two distinct analyzes can be performed. By evaluating the overall precision of evolution of different models, it can be concluded that the linear model has the highest precision in the range 30 to 900 training elements. The logistic model and SVM model presents a homogeneous identical behavior in terms of precision values over the range 150-900 training elements. From the driving elements 900, the noise introduced by the unbalance 2R = NR, causes the linear SVM models and feel abrupt loss of precision.

By evaluating the overall evolution F1 of the different models it can be concluded that the linear model has the highest F1 in the range 30 to 900 training elements. The logistical and SVM models have a similar F1 throughout the range but slightly lower than the linear model. As 2R = NR precision of curve 900 from the driving elements the linear SVM models experienced an abrupt reduction F1 levels due to noise introduced by unbalance 2R = RN. It is worth noting that the logistic model suffered a much steeper fall than the other models,
which means that the logistic model recall was very low, since the precision of the logistic model rose from 900 training elements as referred above.

Figure 7: F1 depending on the number of elements of the training set 2R = NR

5.5.2 ROC curve analysis

This section is intended to compare the ROC curves of the chosen models 2R = NR. As can be seen in Figure 8 all models lies above the line (0,0) a (1,1), this means that outperform the random model. 2 SVM linear models have a similar curvature. However, the linear model is the one closer to the random line. The best classifier is the logistic model. This is easily identifiable graphic view, because this comes close to the point (0,1) which is representative of the perfect model that classifies all relevant training set slide correctly and not relevant.

Anyway it can confirm previous findings using the AUC value associated with each model. The order from best to worst classifier is as follows: \( AUC_{	ext{logistico}} = 0.86 > AUC_{	ext{SVM}} = 0.71 > AUC_{	ext{linear}} = 0.66. \)

6. Global results

Analyzing the situation 1: R = NR, the linear model has the highest precision and F1 when trained with 20 to 80 training set elements. So it is the best model for small datasets. The logistic model has the highest precision, F1 and AUC when compared to other models 80 trained with a training set of 800 elements. So it is the best model for large datasets. Analyzing the situation 2R = NR, The linear model has the highest precision and F1 when trained with training 30 to 90 elements. So it is the best model for small datasets. The linear model has the highest precision and F1 when compared to other models 90 trained with a training set of 900 elements. The logistic model has the highest AUC and a precision and F1 very close to the linear model in this situation.

The situation 2 is more conducive to occur when students are using BiblioMiner because it is expected that the number of relevant driving elements is less than the number of driving elements not relevant (Table 5.1). Starting from the results discussed above, and bearing in mind the class skew property of ROC curve, the system chosen to generate the BiblioMiner the scoring function is the logistic model with Epachnenikov core and shuffled sampling because it is the model that shows the ROC curve more close to the perfect model.

7. Conclusions

The results are promising and can be are subject to improvement. Since this is such a vast area it is possible to complement this work or continue it by
following these development of vectors:
- Addition of new relevant indicators, such as words that are underlined / highlighted in the document that could not be captured due to the existing technology of data mining in pdf’s.
- Create an online business model, in which experts would train datasets applied to the elements of literature students.
- Logistic model had a good performance in this project. It will be interest consider the use of different kernels such as sigmoid.

References


