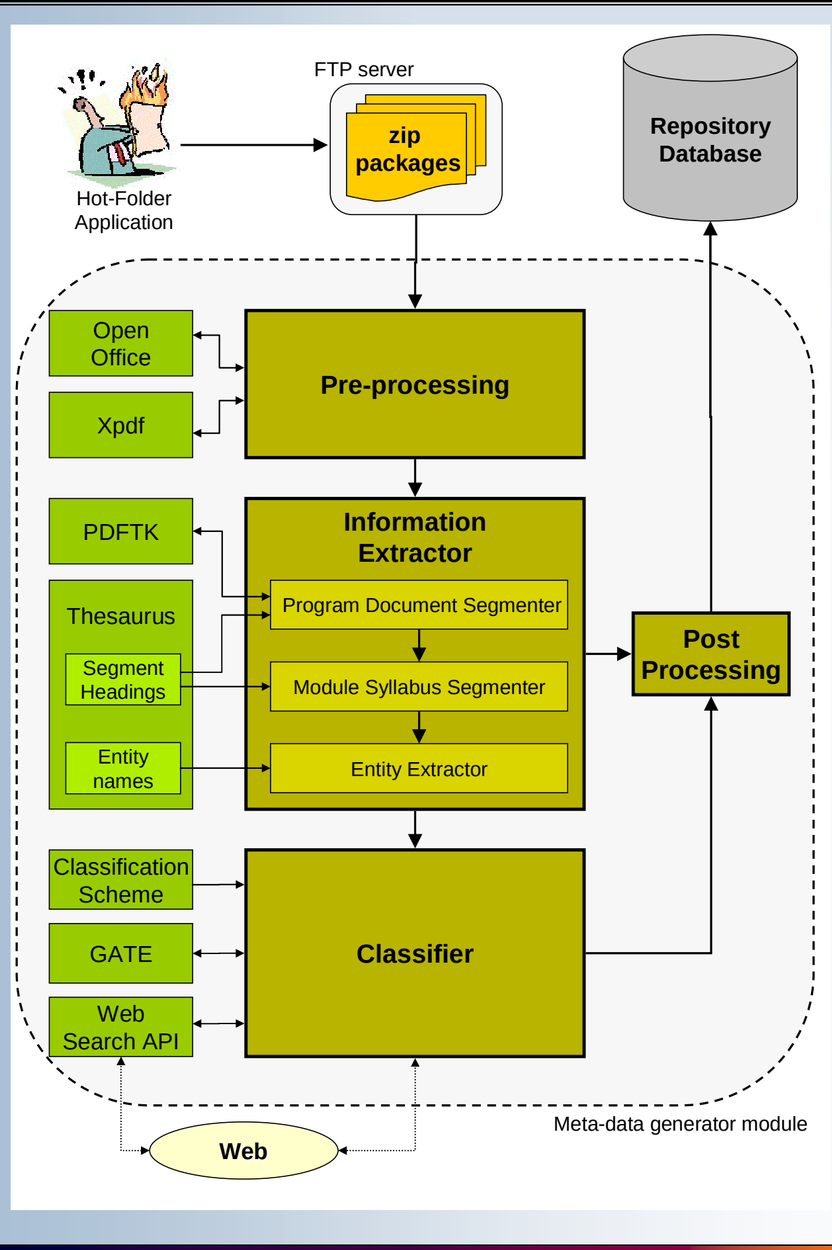
Automatic Classification of Teaching and Learning Materials Based on Standard Education Classification Schemes

Arash Joorabchi and Abdulhussain E. Mahdi
Department of Electronic and Computer Engineering, University of Limerick, Limerick, Ireland
{Arash.Joorabchi, Hussain.Mahdi}@ul.ie

RATIONALE

- Large-scale digital libraries such as the Irish syllabus repository contain thousands of items and therefore require deploying flexible querying and information retrieval techniques that allow users to easily find the items they are looking for.
- Classification of materials in a digital library based on a pre-defined scheme improves the accuracy of information retrieval significantly and allows users to browse the collection by subject.
- Manual classification is a tedious and expensive job requiring an expert cataloger in each knowledge domain represented in the collection and therefore deemed unfeasible in many cases.
- Automated Text Classification/Categorization (ATC) - the automatic assignment of natural language text documents to one or more predefined categories or classes according to their contents - has become one of the key techniques to enhance information retrieval and knowledge management of large digital collections.



SYLLABUS CLASSIFIER

- The task of the syllabus classifier is to automatically assign a classification code to each individual course/module based on a predefined education classification scheme.
- The Higher Education Authority (HEA) and higher educational institutions in Ireland use the International Standard Classification of Education (ISCED) to provide a framework for describing statistical and administrative data on educational activities and attainment in Ireland.
- The need for a more detailed national education classification standard than that provided by the ISCED has already been recognised by educational authorities within other jurisdictions. This has led some other countries to develop their own national classification of education standards such as JACS in UK and ASCED in Australia.
- In order to standardise the classification of modules among Irish higher education institutes, HEA is currently considering the development of Irish Standard Classification of Education.
- The current version of the classifier classifies the syllabus documents based on a draft extended version of ISCED which will be replaced by the Irish Standard Classification of Education in future.
- The syllabus classifier component is based on the widely used Naïve Bayes algorithm. we are also experimenting with the application of a search engine in automatic collection of a training set for creating a fully unsupervised document classification system.

TRAINING THE CLASSIFIER

- A major difficulty of supervised approaches for text classification is that they require a great number of training instances in order to construct an accurate classifier.
- Manual classification of documents for the purpose of training a classifier is a tedious and expensive job. Motivated by this problem, the semi-supervised and unsupervised training methods are being researched to train a classifier with a limited number of training documents and no training documents, respectively
- In this work we experiment with an un-supervised web-based approach to train a Naïve Bayes classifier used for classifying syllabus documents based on a hierarchical education classification scheme.
- The classification scheme used here is an extended version of ISCED represented in XML. The ISCED is a hierarchical scheme with three levels of classification: broad field, narrow field, and detailed filed. Accordingly, the scheme uses a 3-digit code in a hierarchical fashion for classifying fields of education and training , such that the first digit represents ‘broad field’, the second digit represents the ‘narrow field’ and third digit resents the ‘detailed field’ of a given document. There are 9 broad fields, 25 narrow fields and about 80 detailed fields. We have extended this by adding a forth level of classification, subject field, which is represented by a letter in the classification coding system. For example a module assigned the classification code “482B” would indicate that module belongs to the broad field of “Science, Mathematics and Computing”, the narrow field of “Computing”, the detailed field of “Information Systems” and the subject field of “Databases”, where the broad fields, narrow fields and detailed fields represent the branches of the upper three levels of the classification hierarchical tree, from top to bottom respectively, and the subject fields represent the leaves of the tree.

AUTOMATED TRAINING CORPUS COLLECTION

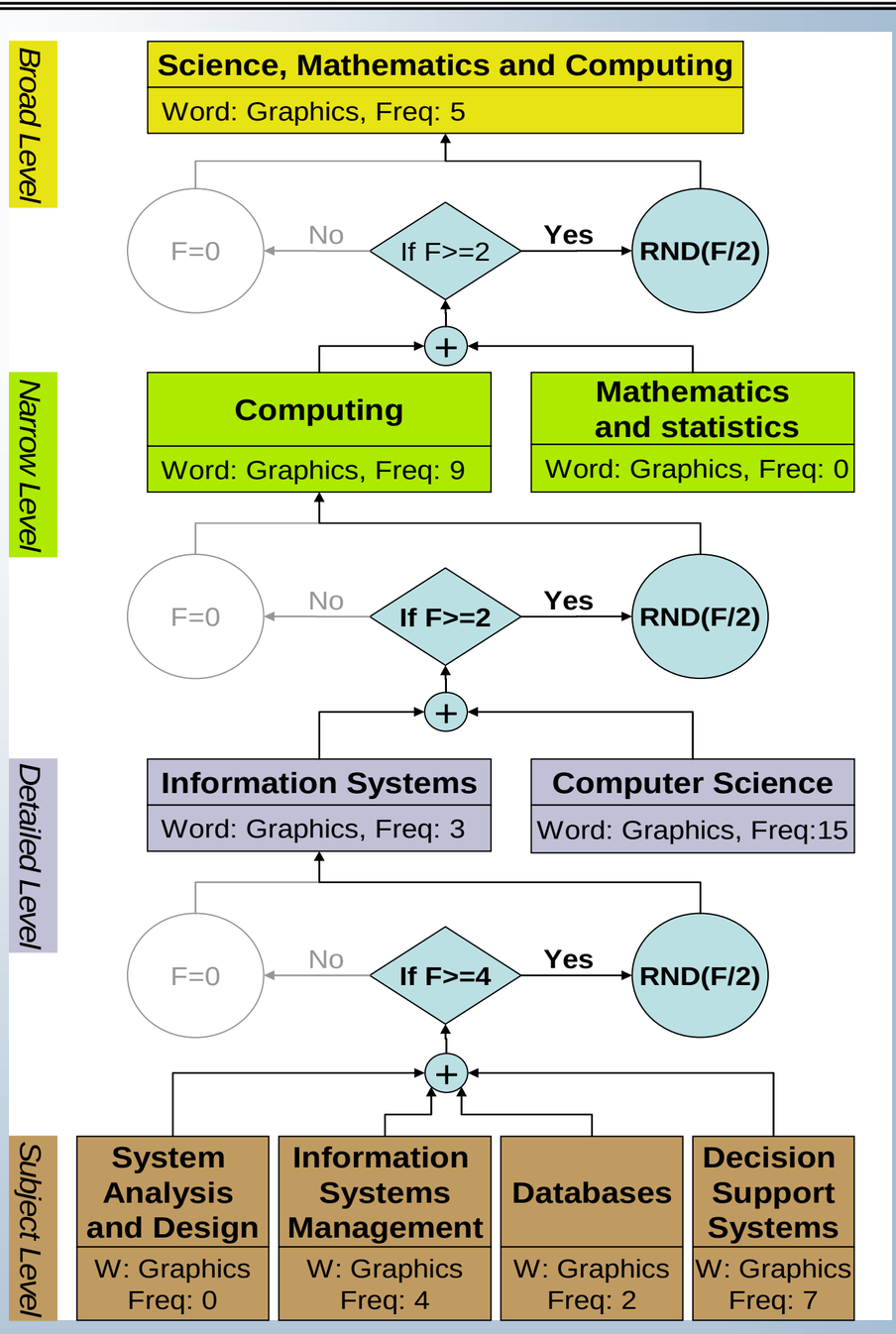
The classifier starts the training process by reading the XML version of classification scheme and collecting a list of subject fields (leaf nodes). Then a search query created from the name of the first subject field in the list combined with the keyword “syllabus” is submitted to the Yahoo search engine using the Yahoo SDK. For example, the query created for the subject filed 482B, databases, is “databases syllabus”. The first hundred URL’s in the returned results are passed to the Gate toolkit, where the files (HTML, Text, PDF, and MS-Word) that they are pointing to are downloaded and their text content is extracted and tokenized. This process is repeated for all the subject fields in the hierarchy. The tokenized text documents resulting from this process are converted to word vectors which are used to train the classifier for classifying syllabus documents in subject-field level and to create word vectors for the fields which belong to the upper three levels of the classification hierarchical tree. The words used in the name of subject fields have direct effect on the quality of search results and using words that have a high information gain value improves the quality of search results however we have not changed the subject field names in this experiment as we wanted to measure the accuracy of the system with a standard classification scheme in its original form. The other factor affecting the quality of search results is the number of learning-teaching documents in each field that are available online. For example the quality of search results for computer related fields such as databases, programming languages, and artificial intelligence is substantially higher than fields such as veterinary nursing or cereal science which are less populated. Our experiment shows that the number of relevant syllabus documents retrieved from the first hundred URL’s of search result can vary between 20 and 40 depending on these two factors. Although the remaining documents are not syllabus documents, the majority of them can be classified to the subject field or its parent (i.e., a detailed field), making them useful for training the classifier. For example looking for syllabus documents in the subject field of databases a lot of the retrieved documents might not be database-related syllabus documents but they still can be classified to the subject field of databases as their main content discusses some aspect of the database systems. Also in majority of cases that the retrieved document can not be classified to the subject field of databases it still can be classified to the detailed field of computer science which is the parent of databases subject field .

The subject-field word vectors created by leveraging a search engine are used in a button-up fashion to construct word vectors for the fields which belong to the higher levels of hierarchy. We illustrate this method with help of the following example. Assume we want to create a vector of words for the detailed field of information systems. There are four subject fields that descend from it: Systems Analysis and Design, Databases, Decision Support Systems, and information systems management.

We build a master vector by combining the vectors of these four subject fields and then normalize the word frequencies by dividing the frequency of each word in the master vector by the total number of subject field vectors used to create it (4 in this case) and then round the quotient to a positive integer number. During the normalization process, if the frequency of a word is less than total number of vectors it will be removed from the vocabulary. The process described above can be expressed as follows:

$$F(w_i) = \begin{cases} 0 & \text{if } \mathrm{Sum}(F(w_i)) < |Fields| \\ \mathrm{RND}\left(\frac{\mathrm{Sum}(F(w_i))}{|Fields|}\right) & \text{if } \mathrm{Sum}(F(w_i)) \geq |Fields| \end{cases} \qquad \mathrm{Sum}(F(w_i)) = \sum_{n=1}^{|Fields|} F(w_{n,i})$$

This process is repeated in a similar fashion to create word vectors for all the detailed, narrow and broad fields of the classification hierarchy in a bottom-up manner. In rare cases where a detailed or narrow field does not have any descendent, the web-based approach is used to create a word vector for higher level fields.



RESULTS

The performance of the classifier was measured using a hundred undergraduate and a hundred postgraduate syllabus documents. The micro-average precision achieved for undergraduate syllabi was 0.75 and it decreased to 0.60 for postgraduate syllabi. Examining syllabus documents from both groups indicates that some syllabi are describing courses which contain components belonging to different fields of study. For example a syllabus document could be describing a course which contains both database design and web design components. Classifying such documents which belong to more than one class is more error-prone and requires the classifier to recognize the core component of the course. Since the number of multi-component courses is substantially higher among the group of postgraduate courses, therefore the classification accuracy achieved for this group of syllabus documents is about 15% lower than undergraduate syllabi. Also it should be noted that this level of accuracy is achieved without using any manually classified training document to train the classifier.