Text analysis module of a System for Automatic eXtraction of lEarning object Features (SAXEF)

Marco Alfano
Anghelos
Centre on Communication Studies
Via Pirandello 40
90144 Palermo, Italy
Tel. +39091341791
marco.alfano@anghelos.org

Biagio Lenzitti
Computer Science
University of Palermo
Via Archirafi 34
90123 Palermo, Italy
Tel. +390916040427
lenzitti@math.unipa.it

Natalina Visalli
SISSIS
University of Palermo
Viale delle Scienze
90100 Palermo, Italy
Tel. +390916040427
vnatalin@neomedia.it

ABSTRACT

New on-line courses are often created by using existing learning objects found on the net. However, those learning objects cannot easily be reused for the creation of a new didactic work because they are usually proposed without information on their aims and the typology of users which they are destined to. Moreover, the contents are not clearly synthesized so that the reading of the whole object is often necessary to understand its relevance to the new course.

To facilitate this task, we have created a system called SAXEF (System for Automatic eXtraction of lEearning object Features) which allows to automatically extract the basic indicators of any learning object (a sort of DNA) found in Internet. It provides a valuable help to a teacher who is in the process of creating a new on-line course because he/she can easily choose the most appropriate learning objects from the net just by looking at their basic indicators. SAXEF presents a modular structure and we have already developed some modules and are in the process of implementing the rest of the system. This paper presents the main architecture of SAXEF and the details of the text analysis module for extracting main and secondary topics of a learning object.

Categories and Subject Descriptors

J [Computer Applications]

General Terms

Design, Experimentation.

Keywords

On line Education, Learning Objects, Internet, Metadata Extraction, Text Analysis.

1. INTRODUCTION

Many on-line "learning objects" (LO) are nowadays available on the net. They satisfy various formation requirements, from the scholastic one (mainly courses aimed to university and post-university formation) to the professional one (basic formation or update courses) and cultural one (courses given by public and private institutes). The proposals can be distinguished for the typology of the content presentation (text, multimedia, etc.), the length and level of details (from the single monothematic lesson to the whole multidisciplinary course), and the interactivity degree (depending upon the interactivity level at user disposal). Moreover, some assume that the student works alone along his/her learning path while others assume an interaction with a tutor (in a synchronous or asynchronous way) [1], [2].

The search of a specific topic in Internet provides a lot of information and much of this information has a didactic structure [3], [4], [5]. This suggests the possibility of their reuse for the creation of a new didactic work [6], [7]. However, the found learning objects cannot easily be reused because they are usually proposed without information on their aims and the typology of users which they are destined to. Moreover, the contents are not clearly synthesized so that the analysis of the whole object is often necessary to understand its relevance to the new course [8], [9].

An help to on-line courses development could come by automatically extracting the main characteristics of existing learning object features for an easier reuse in a new on-line course [10]. The learning objects should be characterized by their contents, communication methodology and required pre-existing knowledge. Moreover, in accordance with the hypertext peculiarity of Internet, they should be linked to each other allowing to retrieve other objects for the full comprehension of the treated subject and its deeper analysis [11], [12]. For example, it would be important to recognize which context a learning object belongs to, evaluate whether its content is either theoretical or practical, synthetic or analytical, to understand what are the main and secondary topics, the level of complexity and the iper/multimedia structure. Such characteristics would also allow to connect those objects to other learning objects [13], [14].

Considering already existing objects, we have thought how to extract characteristics from a complex structure such as that of a learning object without an additional participation of the author who could characterize it through, for example, metadata [15], [16]. This is a fundamental step because we assume that the information obtained by the analysis of the LO components and the study of their relationships allows us to characterize the learning objects through a map, a sort of DNA, that contains the generic and specific elements and totally describes the object.

Starting from this hypothesis, we have developed the architecture of a system called SAXEF [17], [18] (System for Automatic eXtraction of lEearning object Features) which allows to automatically extract the basic indicators of a learning object (the sort of DNA previously discussed) and provides a valuable help to a teacher that has to create a new on-line course.

SAXEF presents a modular architecture and we have already developed some modules and are in the process of implementing the rest of the system. We have also executed some experiments on the system in order to validate the basic hypothesis and prove the usefulness of such a system. The architecture of SAXEF and the implementation details of the developed modules are the subjects of the present paper.

The paper is organized as follows. Chapter 2 describes the SAXEF architecture. Chapter 3 provides the implementation details of the SAXEF module which allows to automatically derive main and secondary topics from a learning object. Chapter 4 provides the results of the first experiments carried on this module and finally, Chapter 5, reports some conclusions and future work.

2. SAFEX: A SYSTEM FOR AUTOMATIC EXTRACTION OF LEARNING OBJECT FEATURES

The SAXEF system has been thought as capable of extracting text/multimedia features of each learning object and the whole course considered as a structured set of learning objects. The structure of the course and that of the learning objects together with the relationship between their media assumes then an important role in determining the nature of the course itself. In practice, given a course or a single learning object, SAXEF will produce an "E-learning Identification Card" (EIC) with the following information on the course/object nature:

- main topics;
- secondary topics;
- theoretical or practical;
- synthetic or analytical;
- media types and multimediality level;
- complexity level;
- links to other EICs with same topics;
- links to other EICs with related topics.

The EICs will be organized in a database and will be shown through a graphical interface indicating the main topic and its connections.

The SAXEF architecture is made up of three levels (Fig. 1):

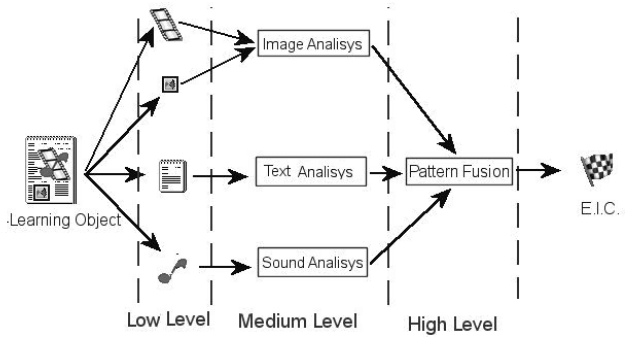

Figure 1. SAXEF architecture.

- a low level to identify and separate all the media components of the learning objects (text, images, video, audio, animations, etc);
- a medium level to extract specific features of each media by using specialized algorithms (text analysis, image analysis, …);
- a high level to fuse the media features and show the structure and the indicators of the learning objects through the creation of their EICs.

It should be noted that the fusion of the elementary data must not be done simply putting together the results of the specific analyses but rather as a further analysis of the complete context. This is done similarly to some algorithms for extracting information from an image where an analysis of the relationship between elements such as vertical and horizontal lines or circular shaped structures is performed.

The availability of EICs then greatly helps either teachers who want to create new on-line courses or students who desire to organize their didactic paths. For example, a teacher who has to organize a new course (on-line or not) on a specific subject and wants to use some existing on-line courses or parts of them, usually uses a common search engine with that subject and obtains a list of web pages that deal with that subject. He/she then analyzes a number of URLs to understand if they have a didactic nature and if the content and the teaching method are suitable to his/her needs. Of course, this process is very time consuming and the teacher will often limit his/her search to the first few URLs. With SAXEF, the teacher can enter whatever number of URLs (of single web pages or whole courses) into the system and SAXEF will provide him/her with the basic indicators described above. From the exam of the few indicators, the teacher will be able to evaluate what are the URLs of interest to him/her and for those URLs, he/she will examine the related web pages for the final analysis and decision whether to include the related contents into the course. It is then clear how SAXEF saves time to the teacher who, on the other hand, keeps full control on the contents of his/her course.

As explained above, the first phase of SAXEF development is devoted to confirm our basic hypothesis, i.e., LO indicators can be derived from information extracted from the various media components and from the study of their relationships. In particular, we have developed the low level architecture to separate the different media of a learning object (given in a standard html format) and a first module of the medium level to perform a text analysis that provides the most frequent words and their distances. The objective is to find main and secondary topics of the learning object or the whole course. This part of SAXEF is available through Internet at the address http://altair.math.unipa.it/saxef.

The next chapters describe the details of the module that implements the text analysis and the results of the first experiments carried on this module to evaluate its precision and usefulness.

3. TEXT ANALISYS MODULE

We have implemented the text analysis module inside SAXEF as a web application using the Perl and PHP languages and a MySQL database. Perl and PHP have been chosen because of their easyness of use, string manipulation capability, optimal web interfacing (HTML and XML) and possibility to insert SQL

queries inside the code. The architecture of this module is shown in Fig. 2.

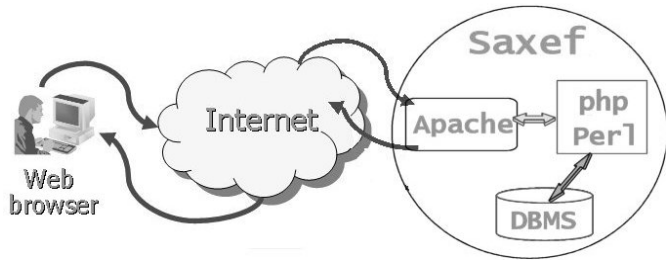

Figure 2. Architecture of the text analysis module.

The module will perform the following tasks:

- Input parameters management
- Link discovery
- Text analysis
- Data insert in the database

We now describe those tasks in more details.

3.1 Input parameters management

Upon starting the module, the user will be presented with a web page that contains a series of fields and checkboxes to be filled with the input data (Fig. 3).

Figure 3. Input web interface.

The user has to insert the URL of the web page to be parsed and decide whether only that page or the whole course has to be parsed. Moreover, he/she can decide the output level (summary or details of the analysis) and the minimal percentage of word occurrence (compared to the total number of words in the text) such as that the word can be considered as a keyword. For example, if the user inserts 3% and the whole text contains 100 words, SAXEF will consider keywords only those words which appear at least three times in the text. This is done for "text normalization", i.e., to avoid that a longer text is considered more meaningful than a shorter text. The percentage can be expressed for single words or couples but its insertion is optional. If not expressed, SAXEF will be able in any case to compute the proper percentages based on the text length.

3.2 Link discovery

The link discovery is activated when the user asks for the analysis of the whole course. The Perl engine will store the main URL (the one inserted by the user) into the database and will pass (through a simple SQL query) this URL to a specific Perl module, called "parse.pm". This module will analyze the web page in search of internal links, i.e., URLs that have the initial part in common with the main URL. Once the internal links are found, they will be stored in the database by the Perl engine and passed to "parse.pm" to find other internal links and so on until all the internal links are found and parsed.

3.3 Text analysis

The text analysis will be executed through the following steps:

- Elimination of the words of common use
- Computation of single word occurrences
- Computation of word couples occurrences
- Research of words inside the relevant tags
- Word selection
- Weight assignment to words
- Choice of main ad secondary topics based on the weighted words

First, all the common words (articles, prepositions, pronouns, common verbs, etc.) have to be eliminated. To this aim, a text file containing the list of those words has been created and this file can be easily modified through the main web interface (Fig. 3).

This file is read when the application is started and the data are stored in a hash table. This table will be examined to evaluate whether a word found in the html page is "valid" or not. This operation is done every time the Perl "analyse" function computes the word occurrences. If the word is "valid", it will be inserted in a hash table and if already present will increment the number of occurrences. The same process is applied to the word couples. In this case both words must be "valid".

The next step is to identify the words inside the "relevant" tags. <title> and <meta> are considered relevant tags. Perl has optimal tools for such a task.

The next step is to consider the most relevant words inside the text. This is achieved by pruning the set of words selected so far taking into account the percentages introduced by the user for occurrences of single words and couples. If the user has inserted x% and y% respectively for the occurrences of single words and word couples and if the number of total words (excluding the common words) is n, then only the single words that appear in the text more than x*n/100 times will be selected and similarly only the y*n/100 couples will be selected. This is done for text "normalization" as explained above and if the user does not fill the percentage fields, these will be automatically filled with the 1/n value. This value has been chosen after running several experiments with different percentages.

The next step, a very important one, is to provide each selected word with a weight. This weight will be used for determining main and secondary topics of the parsed web page. In practice, the weight is a score that the word obtains depending on when and where the word appears in the text.

In order to establish the proper score, we have run some preliminary experiments that have led to the following basic considerations:

- the score (and then the weight) of each word must be a function of the total number of words in the text (again for normalization purposes);

- the single words that also appear in couples must have a greater weight;
- the words that appear inside the <title> tag are particularly relevant and immediately after come the words that are inside the <meta> tag.

Taking into account the above considerations, the weight to each selected word is provided through the following formula:

$$\text{weight} = \text{occ\_w}/\text{tot\_w} + (M/\sqrt{\text{tot\_w}}) + 3*(T/\sqrt{\text{tot\_w}}) + \text{occ\_d}/\sqrt{\text{tot\_w}})$$

where: *occ_w* is the number of occurrences of the single word; *tot_w* is the total number of words; *occ_d* is the number of occurrences of word couples containing the word; *M* is a constant related to the <meta> tag that is applied if the word appears inside the <meta> tag, *T* is a constant related to the <title> tag that is applied if the word appears inside the <title> tag. Note that, in the above formula, T is multiplied by 3 because the <title> tag is considered more important than the <meta> tag.

The final step is to decide the main and secondary topics. This is simply done by considering main topics the words with the two highest scores and secondary topics the words with the following four highest scores.

3.4 Output and data insert in the database

The results of the text analysis together with the related URLs are presented through a web page that contains the main and secondary topics with their scores, and the links of parsed pages (Fig. 4).

Figure 4. Output web page.

The user can delete unrelated topics and links, and store the remaining ones in a MySQL database. Finally any user can perform a search of available topics in the MySQL database. This is easily achieved through simple SQL queries thanks to the possibility to insert those queries inside the Perl and Php code.

4. EXPERIMENTAL RESULTS

After the development of the text analysis module, we have led a series of informal tests for a first evaluation of results precision. The main objective of these tests has been to evaluate the general correctness of the method and the related choices.

The methodology used for the tests has been the following:

– Search of a topic in the net by means of a search engine (Google).

– Choice of some learning objects found in the net.

– Human analysis of chosen learning objects for determining the main topic.

– Text analysis of same learning objects through SAXEF.

– Score attribution to words found by SAXEF. We have given 1 point to first and second found word (main topics) and 0.5 points to the next four words (secondary topics) when we found a thematic correspondence between the word and the main topic determined through the human analysis.

– Computation of the proximity index as a sum of the individual scores. The proximity index can then vary from 0 to 4.

The table below shows the applied methodology for a specific URL.

Table 1. Example of test table.

URL	http://www.synapses.co.uk/genetics/index.html
Human Analysis	Genetic
SAXEF Analysis and score attribution	1. chromosomes (1) 2. cell (1) 3. meiosis (0.5) 4. genetics (0.5) 5. mitosis (0.5) 6. number (0)
Proximity index	3.5

We ran fifty tests using heterogeneous topics (from history, arts, mathematics, physics, and so on) and choosing whole courses with at least three learning objects.

Fig. 5 reports the results of those tests expressed as a distribution of the proximity index. Note that almost 90% of the proximity indexes are greater than the mean value (2) and that 70% of the proximity indexes are greater than or equal to 3 indicating that at least four words are related to the theme treated by the chosen topic.

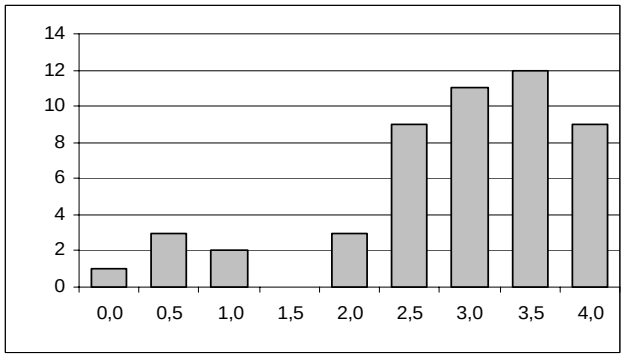

Figure 5 Test results.

We understand that this analysis must be more accurate and less linked to human judgments and we will continue the tests refining the experimental methodology. Nevertheless, these first results are very encouraging and suggest us to progress further in this direction.

5. CONCLUSIONS AND FUTURE WORK

This paper has dealt with the possibility of creating new on-line courses by using already existing learning objects or whole on-line courses. This is achieved through automatic extraction of learning object features and creation of a specific E-Learning Identification Card (EIC) for each object. We are developing a system (SAXEF) for creation and storage of such EICs. SAXEF will allow teachers to easily find learning objects with desired contents and structure to create new on-line courses.

SAXEF development is planned in more steps and the first phase has been mainly devoted to dividing a learning object in its media components and performing a textual analysis to find main and secondary topics. Presently only html documents that end with .htm or .html can be analyzed through SAXEF but soon also html documents ending with .php, .js and .asp will be permitted together with doc, pdf and ppt documents. Moreover, a second module is being developed to analyse other media in order to extract further information on the LO content and to understand whether the LO is either synthetic or analytical and its multimediality level. Finally, a graphical map of all available learning objects/courses is being created. The map will represent each learning object by means of its main topic(s) and will graphically show the relationships with the other learning objects by means of lines with varying distances. This will allow teachers to have a synthetic view of the available topics, the related learning objects and the connections among those objects.

6. REFERENCES

[1] Lenzitti, B., and Visalli, N. STDIO.D Strumenti per la Didattica On line.Docenti. In *Proceedings of Expo-Learning 2004.* October 9-12, 2004.

[2] Alfano, M., Lenzitti, B., and Pace, A.. Tutor-Sky: A web environment for multimedia on-line education. *Methodologies and Technologies for Learning.* G. Chiazzese et al. (Eds.), WIT Press, 2005, 297-304.

[3] Martinez, M.. Designing learning objects to mass customize and personalize learning. *The Instructional Use of Learning Objects.* D. A. Wiley (Ed.), 2000.

[4] Koper, R., and Tattersall, C. *Learning Design. A Handbook on Modeling and Delivering Networked education and Training*. Springer, Berlin, 2005.

[5] Calvani,, A.., and Rotta M. *Fare formazione in Internet. Manuale di didattica ondine*. Erickson, 2000.

[6] Hodgins, H. W. The future of learning objects. *The Instructional Use of Learning objects.* D. A. Wiley (Ed.), 2000.

[7] Collins, B., and Strijker. A. New Pedagogies and re-usable learning objects; toward a new economy in education. *Educational Technology Systems*, vol. 30(2), 2001-2002 137-157.

[8] Williams, D. D. Evaluation of learning objects and instruction using learning objects. *The Instructional Use of Learning objects.* D. A. Wiley, 2000.

[9] Costa, R., and Galiani, L. *Valutare l'e-learning.* Pensa, 2003.

[10] Petrucco, C. Le Prospettive Didattiche del Semantic Web. In *Proceedings of Didamatica 2003*. Februray 27-28, 2003, 168 -176.

[11] Alvino S., and Sarti L.. Learning Objects e Costruttivismo. In Proceedings of *Didamatica 2004.* Ferrara, May 10-12, 2004.

[12] Gibbons, A. S., Nelson, J., and Richards, R. The nature and origin of instructional objects. *The Instructional Use of Learning Objects.* D. A. Wiley, 2000.

[13] Wiley, D. A. Connecting learning objects to instructional design theory: a definition, a metaphor, and a taxonomy. *The Instructional Use of Learning Objects.* D. A. Wiley, 2000.

[14] Merrill, M. D. Knowledge objects and mental models. *The Instructional Use of Learning Objects.* D. A. Wiley, 2000.

[15] IMS Global Learning Consortium. *Learning Resource Meta-data Specification*. http://www.imsglobal.org/metadata/

[16] IEEE-SA Standards Departemt. *Draft Standard for Learning object Metadata*. 2002

[17] Alfano, M., Lenzitti, B., and Visalli, N. Creation of on-line courses using existing learning objects. In *Proceeding of E-Learning Conference 2005.* Berlin, September 6-7, 2005.

[18] Alfano, M., Lenzitti, B., and Visalli, N. SAXFE: A System for Automatic eXtraction of learning object Features. In Proceedings of II Sie-L Conference. Florence, November 9-11, 2005.