SAXEF: A System for Automatic eXtraction of E-learning object Features

Marco Alfano[a], Biagio Lenzitti[b], Natalia Visalli[c]

[a]Anghelos Centre on Communication Studies; [b]Dept. of Mathematics - Un. Palermo; [c]SISSIS - Un. Palermo; Italy

[a]marco.alfano@anghelos.org; [b]lenzitti @math.unipa.it; [c]vnatalin@neomedia.it

Key words: E-learning, Learning Objects, Didactic Indicators, Metadata Extraction.

Abstract

New online courses are often created by using existing materials on the net. However, those materials 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 a complete analysis of the whole materials is often necessary to understand their relevance to the new course. Using our experience on the creation of online courses with existing web materials, we have thought how to help teachers in finding the best materials for the creation of new online courses. To this end, we have developed a system, called SAXEF (System for Automatic eXtraction of lEearning object Features), that is capable to automatically extract the didactic indicators (a sort of DNA) of any web page (or group of pages) found on internet and allows a teacher to easily evaluate whether that page (with its contents) is of interest to him/her. This paper presents the main architecture of SAXEF, its implementation details and some experiences on its use. At present, SAXEF is capable to automatically extract various didactic indicators such as main and secondary topics, synthesis level and multimedia level from any group of web pages.

1 Introduction

There is an increasing number of learning materials on the net that satisfy various training requirements, from the scholastic one (mainly courses aimed to university and post-university training) to the professional one (basic training or refresher 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 on the interactivity level at disposal of the user). Moreover, some courses assume that the student stays alone along his/her learning path while others assume an interaction with a tutor in a synchronous or asynchronous way (Alfano *et al.*, 2005; Calvani and Rotta, 2000; Lenzitti and Visalli, 2004).

Since its birth, internet has been used as a learning tool thanks to the amount and variety of information at hand. Nowadays, it is even more used as much of the information in the net has a didactic nature and many suitable platforms for e-learning exist. The creation of didactic material for the net seems however to be mainly perceived as a private elaboration aimed to a group of specific users. Nevertheless, publishing information on the net usually means offering it to a much wider audience.

The search of a specific topic on internet provides a lot of information and much of this information has a didactic structure (Calvani and Rotta, op.cit.; Koper and Tattersall, 2005). This suggests the possibility of its reuse for the creation of a new didactic work (Collins and Strijker, 2001; Hodgins, 2000). However, the found materials cannot immediately 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 materials is often necessary to understand their relevance. We wonder whether the search for didactic information on internet can be simplified, so that an optimal use of existing resources can be achieved.

The aim of this work is to present the details of a system, SAXEF (Alfano *et al.*, 2005bis; Alfano *et al.*, 2006), that allows to automatically extract the didactic indicators (a sort of DNA) of any web page (or group of pages) and helps a teacher to easily evaluate whether that page (with its contents) can be useful for a new online course.

The paper is organized as follows. Chapter 2 describes an experience on building an online course with internet materials and the related considerations. Chapters 3 and 4 show the architecture and the implementation details of the SAXEF system. The final chapter describes some experiences with SAXEF together with the conclusions and future work.

2 Building an on-line course with existing materials: Experiences and considerations

Our work started with the development of an online course on "information and communication technologies for didactics" at the SISSIS School (university post-graduate course for school-teachers) of the University of Palermo. Since the goal was to introduce future teachers to the techniques of online education, it seemed natural to make the course materials accessible through the net. Moreover, internet is the best place where to gather such materials.

Using search engines to scan internet, we found several pages with a rich and clear content for each topic. Then, the most suitable pages were chosen considering the aims of the course and the skills of the students. The next step would have been summarizing and re-elaborating the contents of the chosen internet pages, but, in this way, we would only supply a further version of already existing information without adding anything new. On the contrary, we chose to create some specific web pages that contained the links to the internet pages not as further reading but as the lesson contents. The course has then been organized as an "ad hoc" filter to internet resources connected through a graphical diagram[1].

The students experience has been positive because of the accessibility to various documents with basic and advanced information and the possibility of creating individual learning paths. The course is in fact in the net and not simply put on the net (Lenzitti and Visalli, op.cit.).

Another advantage of the internet pages seen as learning objects, has been the possibility of reusing them in other courses with different objectives and audiences. It has allowed us not only to reuse the online contents of other authors but also to capitalize our research work and creation of filter pages. This has been possible thanks to the analysis made on the original objects that has provided us with a complete knowledge of the characteristics of each object, not only on its contents but also on its aims and used language.

Overall, this experience, although positive, has been very time consuming and has taught us that the search for didactic materials on the net is often complex because there is no homogeneous indexing of objects and their characteristics cannot be determined without an in-depth analysis. Moreover, information in internet grows very rapidly and data are introduced from several points in a disjoint way without taking into account what is already present.

Nevertheless, we believe that internet is the best place where to find didactic materials because almost any web page has a didactic potential. We also believe that a teacher should be helped in finding easily and rapidly the materials that are suitable to his/her didactic needs.

An help to online courses development may come from the knowledge of the

1 The course is available at the address http://sissis.unipa.it/sito/levis.

main characteristics of the examined materials without the need of a complete analysis (Fini and Vanni, 2004; Petrucco, 2003). Those materials considered as 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 (Alvino and Sarti 2004; Gibbons *et al*. 2000).

This sort of information is usually contained in the learning objects metadata (LOM) that follow such standards as IEEE and IMS (IEEE, 2002; IMS, 2006). Unfortunately, not all authors are willing to insert metadata when creating new learning objects and, even worse, there is already a huge amount of didactic information on internet that is not structured in the form of standard learning objects and does not contain any metadata.

To overcome this limitation, we have chosen to consider the whole internet as the repository where to take the didactic material from and we have thought how to extract the didactic characteristics of any web page without the presence of additional information such as metadata. This is a fundamental step because we make the hypothesis that the analysis of the web-page components and the study of their relations can provide us with a sort of DNA of that page that contains its basic characteristics including the didactic information.

We then consider any single web page or group of web pages respectively as a learning object (in its broadest meaning) or as an online course and analyze them to extract the main characteristics. In particular, we have worked to recognize which context a web page (or group of pages) belongs to, evaluate whether its content is theoretical or practical, synthetic or analytical, to understand what are the main and secondary topics, the level of complexity and the iper/multimedia structure. This has brought us to the creation of the SAXEF system that will be described in the next chapters.

Other researchers have considered the difficulty of inserting metadata in learning objects and They have tried to automatically extract them. Their analysis is however mainly focused on text (Saini and Ronchetti, 2003; Sonntag, 2004) whereas we also analyze the multimedia contents. Moreover, they extract some of the standard metadata (Brooks, 2006; Cardinaels, 2005) whereas we also define new didactic indicators.

3 SAXEF: System for Automatic eXtraction of E-learning object Features

The SAXEF system has been thought as capable of extracting text/multimedia features from each web page (considered as a learning object in the terms indicated above) or a group of web pages (which represents a whole course).

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 produces 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 are organized in a database and are shown through a graphical interface indicating the main topics and their connections.

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

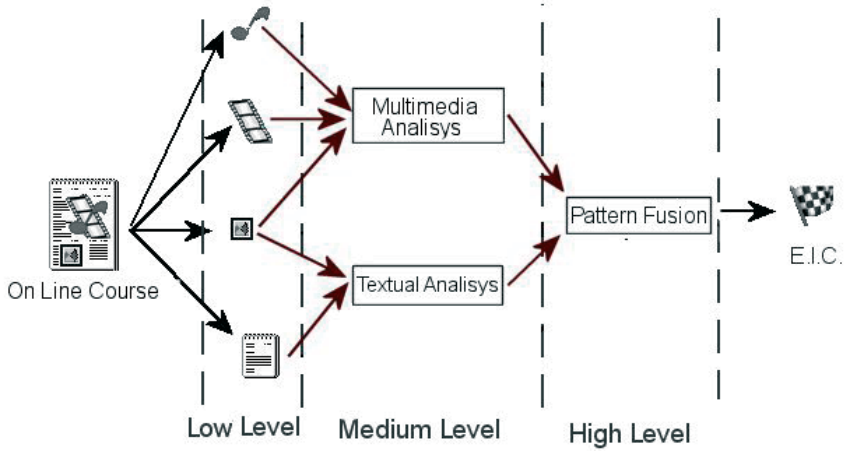

Fig. 1. SAXEF architecture.

1. a *low level* to identify and separate all the media components of the learning objects (text, images, video, audio, animations, etc);
2. a *medium level* to extract specific features of each media by using specialized algorithms (text analysis, multimedia analysis, …);
3. 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 that

extract information from an image where an analysis of the relation between elements such as vertical and horizontal lines or circular-shaped structures is performed.

SAXEF greatly helps either teachers who want to create new on-line courses or students who desire to organize their didactic paths. The user, in fact, 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 user will be able to evaluate what are the URLs of interest to him/her for the final analysis and decision whether to include the pages into a new course. It is then clear how SAXEF saves time to the user who, on the other hand, keeps the full control of the contents of his/her course.

4 SAXEF implementation

We have implemented the low level analysis to separate the different media of a web page, text and multimedia analyses at the medium level and pattern fusion at the high level. In doing so, we have been capable to extract most of the EIC indicators listed above[2].

SAXEF has been implemented as a web application using the Perl and PHP languages and a MySQL database. Perl and PHP have been chosen because of their usage easiness, string manipulation capability, optimal web interfacing (HTML and XML) and possibility to insert SQL queries inside the code. The implementation scheme is shown in Fig. 2.

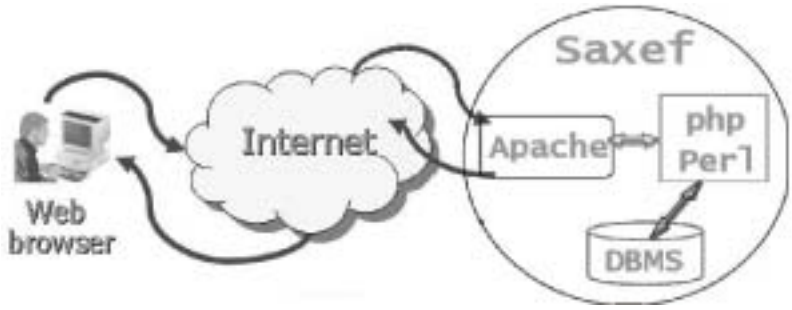

Fig. 2. SAXEF implementation.

4.1 Low-level analysis

After the user provides the address of the web page (or whole course) to be examined, SAXEF takes this page and analyzes its code (in html, xhtml, asp, php, etc.).

2 SAXEF can be found at the address http://altair.math.unipa.it/saxsearch.

SAXEF finds the different objects composing the page and stores the related paths in the database together with the address of the web page. If the user has chosen to examine the whole course, all the pages that are referred to from the main page and have the same root url will be analyzed. This process will proceed recursively until all course pages are analyzed.

4.2 Medium-level text analysis

The text analysis is executed on the text part of the web page through the following steps:

1. All the common words (articles, prepositions, pronouns, common verbs, etc.) are 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;
2. single words occurrences and word couples occurrences are computed;
3. words inside the "relevant" <title> and <meta> tags are identified;
4. the most relevant words inside the text are found. This is achieved by pruning the set of words selected so far and considering specific percentages for occurrences of single words and couples;
5. each selected word is provided with a weight. In practice, the weight is a score that the word obtains depending on when and where the word appears in the text.

4.3 Medium-level multimedia analysis

For each web page the textual and multimedia areas are computed. The textual area is determined by multiplying the number of characters of the web page by the area occupied by each character. We estimate that each character of average size occupies an area of about 100 pixels. We choose an average size for each character because the information on the character size is not always present in the page code (e.g., such information might be contained in external style sheets). The multimedia area is determined by summing up the areas of the multimedia objects present in the web page. In particular, we consider the sizes (in pixels) of images, videos and animations. Moreover, if and audio file is present, we consider its size (in bits) and divide it by 16 bits (sampling size).

4.4 High-level analysis

The high-level analysis of SAXEF is capable of computing the following EIC indicators:

- Main ad secondary topics.

This is done by taking the results of the text analysis and considering main topics the words with the two highest scores and secondary topics the words with the following four higher scores.

- Synthetic or analytical level.

Assuming that the area of the web page is the sum of the textual area and the multimedia area, the ratio between the textual area and the total area will provide the analytical index (expressed as a percentage). At the same time, the ratio between the multimedia area and the total area will provide the complementary synthetic index.

- Media types and multimediality level.

The multimediality level (expressed as a percentage) indicates the presence of the different media types in a web page (or course). The multimediality index is computed as follows:

- if only text is present, the index is equal to 20%;
- if images are present, the index will be increased of a percentage between 0 and 20% proportional to the image area. For areas greater than 20000 pixels the percentage will remain equal to 20%;
- if audios are present, the index will be increased of 20% only if no video files are present;
- if videos are present, the index will be increased of 40%;
- if flash or other types of animations are present, the index will be increased of 20%.

The index will be equal to 100% when all the media types are present.

5 Conclusions and future work

This paper has dealt with the possibility of creating new online courses by using existing internet materials. This is achieved through automatic extraction of didactic indicators and creation of a specific E-Learning Identification Card (EIC) for each analyzed web page. We have developed a system, SAXEF, for the creation and storage of such EICs. SAXEF allows users to easily find web pages or group of pages with desired didactic contents and structure.

SAXEF presents a modular structure and most of its modules have already been implemented. This modularity also provide us with the possibility to perform separate analyses. In particular, the text and multimedia analyses are executed by two independent web applications which produce their own outputs and tables.

We have run a set of experiments on the two applications to verify their usage easiness and results accuracy. We have seen that the text analyzer is quite efficient and provides results similar to the ones obtained through human analysis (Alfano *et al.*, 2006). On the other hand, the results of the more com-

plex multimedia analysis are quite complete but can be synthesized with more difficulty. Moreover, the potential presence of multimedia elements unrelated to the page contents (e.g., banners) can alter the results of the automatic analysis. To overcome this sort of inconveniences, both applications provide the user with the possibility to eliminate some of the results (in terms of found words or multimedia elements) and then re-calculate the EIC indicators.

At present, SAXEF is providing quite stable results but we are currently running further tests and refining the text and multimedia modules based on the test results. Moreover, we are looking for other didactic indicators of interest to the user (to be added into the EIC) and are devising the related analyses to extract them. Finally an e-learning search engine is being built around SAXEF. It will allow the user to make requests in terms of didactic indicators and will automatically browse the internet to find the web pages that best match the user requirements.

BIBLIOGRAPHY

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

Alfano M., Lenzitti B. & Visalli N. (2005), *Creation of on-line courses using existing learning objects*, in: Proceedings of II E-Learning Conference. Berlin, 6-7 September 2005.

Alfano M., Lenzitti B. & Visalli N. (2006), *Text analysis module of a System for Automatic eXtraction of lEarning object Features (SAXEF)*, in: Proceedings of III E-Learning Conference. Coimbra, 7-8 September 2006.

Alvino S., Sarti L. (2004), *Learning Objects e Costruttivismo*, in: Andronico A., Frignani T., Poletti G. (eds), Proceedings of Didamatica 2004. Ferrara, 10-12 May 2004.

Brooks C. et al. (2006), *Issues and Directions with Educational Metadata*, URL: http://pami.uwaterloo.ca/pub/hammouda/i2lor06-metadata.pdf (accessed on 30th March 2007).

Calvani A., Rotta M. (2000), *Fare formazione in Internet. Manuale di didattica online*, Erickson.

Cardinaels K., Meire M. & Duval E. (2005), *Automating Metadata Generation: the Simple Indexing Interface*, in: Proceedings of the 14th ACM International Conference on World Wide Web, Chiba, 14-15 May 2005.

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

Fini A., Vanni L. (2004), *Learning Object e metadati. Quando, come e perchè avvalersene*, Trento, Erickson.

Gibbons A. S., Nelson J. & Richards R. (2000), *The nature and origin of instructional objects*, in: Wiley D.A. (ed.), *The Instructional Use of Learning objects*.

Hodgins H. W. (2000), *The future of learning objects*, in: Wiley D.A. (ed.), *The Instructional Use of Learning objects*.

IEEE Learning Technology Standards Committee (2002), *IEEE Standard for Learning Object Metadata*, 1484.12.1-2002.

IMS Global Learning Consortium (2006), *IMS Learning Resource Meta-data Specification v. 1.3*.

Koper R., Tattersall C (2005), *Learning Design. A Handbook on Modelling and Delivering Networked education and Training*, Berlin, Springer.

Lenzitti B., Visalli N. (2003), *STDIO.D Strumenti per la Didattica On line.Docenti*, in: Expo-Learning 2004, Ferrara, 9-12 October 2004, URL: http://sissis.unipa.it/sito/levis/articolonew (accessed on 30th March 2007).

Petrucco C. (2003), *Le Prospettive Didattiche del Semantic Web*, in: Proceedings of Didamatica 2003. 168-176, TED, 27-28 February 2003.

Saini P., Ronchetti M. (2003), *Semantic Based Architecture for E-Learning*, Journal of Digital Contents 2 (1), 26-30.

Sonntag, M. (2004), *Metadata in E-Learning Applications: Automatic Extraction and Reuse*, in: Hofer, C., Chroust, G. (eds), IDIMT-2004, 12th Interdisciplinary Information Management Talks. 219-231, Linz, 2004.