Searching the internet for learning materials through didactic indicators

Marco Alfano, Biagio Lenzitti

Abstract: *Internet offers a huge amount of didactic materials that can be used in creating new online courses. However, those materials need a deep analysis to understand their context and contents before their potential use. As a consequence, the search of didactic material in internet is often quite tedious and time consuming so the searcher usually limits his/her analysis to the first found web pages .*
To help users in finding efficiently and timely the most appropriate online materials, we have developed a system, called SAXEF (System for Automatic eXtraction of IEearning 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. Moreover, we have developed an e-learning search engine, SaxSearch, around SAXEF. It allows the user to make requests in terms of didactic indicators and automatically browses the internet to find the web pages that best match the user requirements.

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

INTRODUCTION

The search of a specific topic on internet provides a lot of information and much of this information has a didactic structure [5], [13]. This suggests the possibility of its reuse for the creation of a new didactic work [7], [10]. 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.

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 main characteristics of the examined materials without the need of a complete analysis [8], [14]. 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 understanding of the treated subject and its deeper analysis [3], [9].

This sort of information is usually contained in the learning objects metadata (LOM) that follow such standards as those of IEEE and IMS [11], [12]. 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 consider the whole internet as the repository where to take the didactic material from and we extract the didactic characteristics of any web page without the presence of additional information such as metadata. This is a fundamental step because we assume 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 synthetic or analytical, to understand what are the main and secondary topics, the level of complexity

and the multimediality level. This has brought us to the creation of SAXEF [1], [2], a system that allows to automatically extract the didactic indicators and constitutes the basis of a search engine, SaxSearch, that can be used by any user to find didactic materials in internet through the specification of his/her didactic needs.

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 [15], [16] whereas we also analyze the multimedia contents. Moreover, they extract some of the standard metadata [4], [6] whereas we also define new didactic indicators.

The paper is organized as follows. The second and third chapters show the architecture and the implementation details of the SAXEF system. The fourth chapter describes the working principles of the SaxSearch engine. The final chapter describes the first experiences with SaxSearch together with some conclusions and future work**.**

SAXEF: A 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):

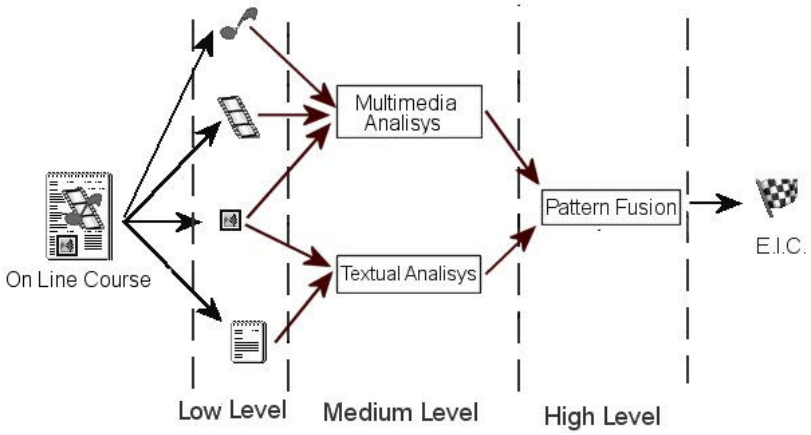

Fig. 1. SAXEF architecture.

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 relationship between elements such as vertical and horizontal lines or circular-shaped structures is performed.

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.

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.

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.).

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.

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.

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., it 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).

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.

SAXSEARCH: AN E-LEARNING SEARCH ENGINE BUILT AROUND SAXEF

Starting from the SAXEF system, we have decided to build an e-learning search engine, SaxSearch, that allows the user to make requests in terms of didactic indicators and automatically browses the internet to find the web pages that best match the user requirements. Fig. 2 shows the implementation structure of SaxSearch. It is implemented as a web application and makes use of a search engine (Google in our case) to find web pages of potential interest. It then uses SAXEF for a complete analysis, extraction of indicators and final comparison with the user requirements.

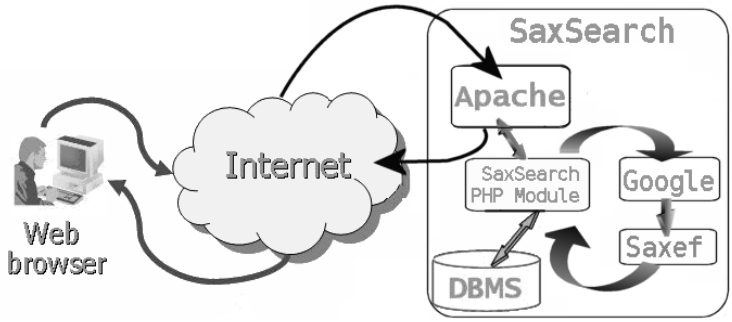

Fig. 2. SaxSearch implementation scheme.

Fig 3 shows the input interface of SaxSearch[1]. The user can specify the topics to search (corresponding to the main and secondary topics shown above), the number of found results per page (in the google style), whether to analyze a single web page or the

[1] SaxSearch can be found at the address http://altair.math.unipa.it/saxsearch.

whole course, the didactic style (syntetic, medium, analytical), the multimediality level (low, medium, high), the language used in the web page, and the date of creation (only pages modified after a certain date will be selected).

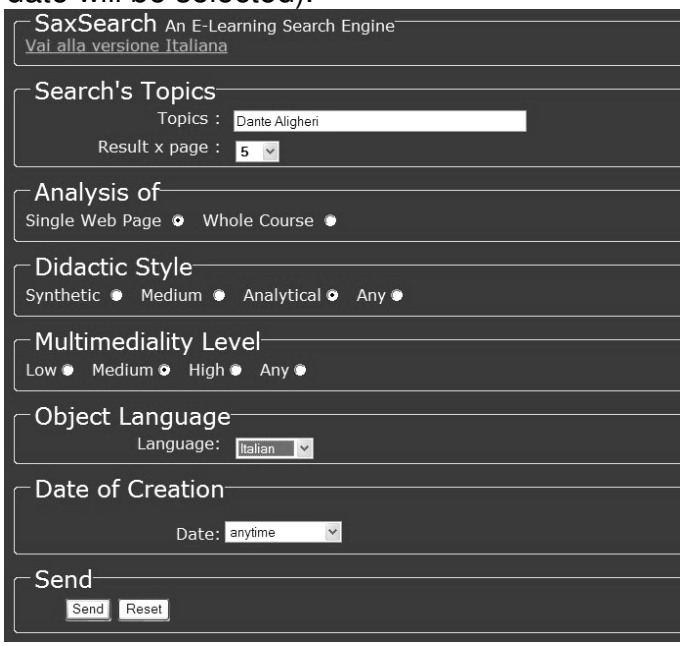

Fig. 3. SaxSearch input interface.

Upon a user request, SaxSearch will produce a list of suitable web pages and their correspondence to the didactic indicators as obtained through the SAXEF analysis.

Fig 4 shows a list of pages that deal with "Dante Alighieri" and have "analytical" didactic style, "medium" multimediality level, "Italian" language and no specific date of creation. The fifth result matches most of the user requirements. By clicking on the results of the text and multimedia analyses, the specific details of those analyses are provided.

Fig. 4. SaxSearch output.

CONCLUSIONS AND FUTURE WORK

SaxSearch 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. On the other hand,

the results of the more complex 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 problems, both applications provide the user with the possibility to eliminate some of the results (in terms of found words or multimedia elements) and then recompute the EIC indicators.

At present, SaxSearch provides 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.

REFERENCES

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

[2] Alfano M., Lenzitti B., Visalli N., 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.

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

[4] Brooks C. et al. Issues and Directions with Educational Metadata, URL: http://pami.uwaterloo.ca/pub/hammouda/i2lor06-metadata.pdf, 2006.

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

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

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

[8] Fini A., Vanni L., Learning Object e metadati. Quando, come e perchè avvalersene, Trento, Erickson, 2004.

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

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

[11] IEEE Learning Technology Standards Committee. IEEE Standard for Learning Object Metadata, 1484.12.1- 2002.

[12] IMS Global Learning Consortium, IMS Learning Resource Meta-data Specification v. 1.3.

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

[14] Petrucco C., Le Prospettive Didattiche del Semantic Web, in: Proceedings of Didamatica 2003. 168-176, TED, 27-28 February, 2003.

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

[16] Sonntag, M., 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.

ABOUT THE AUTHORS

Prof. Marco Alfano, PhD, Anghelos Centre on Communication Studies, Palermo Italy, Phone: +39 091 341791, E-mail: marco.alfano@anghelos.org

Prof. Biagio Lenzitti, Researcher, Dipartimento di Matematica ed Applicazioni, University of Palermo, Phone: +39 091 6040427, E-mail: lenzitti@math.unipa.it