# Automatic Retrieval of Complementary Learning Material for Slide Presentations

*Valdemar Graciano Neto[1], Luciana Oliveira e Silva[2], Ana Paula Ambrósio[3]*

Federal University of Goiás

## Abstract:

*Teachers often use slides to summarize class content and then make them available, through virtual environments, so students can use them as a study reference. Usually these slides do not cover all the content the student should study and the use of complementary material is desirable. It is important that students have access to quality material that will help increase their knowledge, critical sense and coverage about that theme. Especially in introductory courses, students do not have the maturity to select adequate material, and it is up to the teacher to select complementary material and make it available to the students, often in the same virtual environment. This material must be constantly updated by the teacher, that is responsible for searching for new material and associating it to the course content. This paper proposes the use of a Slide Information Retrieval System to automate the association of complementary learning material to slide presentations providing updated material and acquitting the teacher of the research task, ensuring material quality and concordance.*

**Key words:** *slide presentation, automatic indexing, study material.*

# 1 Introduction

Every time a teacher becomes responsible for a course, he usually starts the course definition process by looking for material related to the subject he is going to teach. The search for books, slides, digital material (as pdfs, docs, images), etc. can be useful to improve the teacher's knowledge about that subject and help students acquire the necessary information to complement their studies. The teacher often provides the student with sources that will help him expand his understanding of the subject and help him to complete the tasks related to that subject, such as exercises and surveys.

Virtual Environments have been suitable tools to support Distance Learning in many contexts, from universities to organizations. Attractive features as multi-sync allow learners to access them at any time, ensuring learner's independence and allowing them to keep studying and working anywhere there is a computer connected to the Internet. It has therefore become usual practice among teachers the use of these environments not only for distance education but as a course managing resource for on campus courses as well. The possibility of posting material and information for students has attracted not only the teachers, but the learners as well.

The use of adequate and easily accessible teaching materials (books, handouts, scientific articles, e-books, Web pages, among others) is crucial to guide and motivate the student, and broaden their knowledge on the subject, rising it to a critical level of content understanding, instigating the student to research, while monitoring and evaluating the learning process [1].

The identification of a document's context is carried out by human beings in a natural way. The simple reading of a heading or a summary, for example, can elucidate the content of the material at hand. When carried out by an information system, this task is no longer trivial. This happens because computers do not have the same interpretation capacity as human beings. For this reason, the use of techniques that identify a document's context has been a challenge for researchers for many years.

Search Engines are useful mechanisms to help find relevant material in the Internet where it is possible to find all sorts of material about some topic using a set of keywords. Nevertheless, recovering relevant documents can be a tiring and time consuming exercise. Since human language is naturally ambiguous and polysemous, specifying a query can turn out to be a complex task. Ambiguous queries may retrieve useful as well as useless material that must be analysed and separated to be of use.

On the other hand, teachers are still strongly attached to the use of slides as a presentation tool. The incorporation of new technologies into slide presentations includes the use of ink for annotation and capabilities that allow student-teacher communication [3]. This work aims to further extend a slide presentation tool with contextualized document search capabilities, offering support to students in the process of self-learning, especially at times when additional literature is required. At this point the tool aims to significantly reduce the time spent in the research process by carrying out a search with automatic contextualization, and maximize the recovery of relevant documents.

This paper proposes the use of the Slide Information Retrieval System (SIRS) to automate the retrieval of course related material in the Internet, saving teacher and students of the search cost - that includes problems caused by ambiguity and polysemy – providing updated content about the course theme.

When opening a slide using SIRS, information recovery techniques, applied automatically to the document, identify the words considered relevant in a slide and converts them into hyperlinks, allowing students access to complementary didactic material, either previously selected by the teacher and stored in a repository or retrieved from the Internet. For document searching on the Internet, we foresee a differentiated treatment, whereby searching is performed in academic repositories, and not only through keywords and expansions.

# 2 Information Retrieval

The web is the richest information repository of human history [4]. However, most of the data in the Internet is not structured. Information Retrieval techniques have been developed to help deal with this problem [5]. Their purpose is to find documents whose contents are as close as possible to the user's intentions. However, some Information Retrieval classic problems still affect the result's quality. Such problems include:

1) synonymy - authors and users may sometimes use different words to describe the same concept [6].
2) semantic ambiguity [7] - under a same word, the user can find relevant and irrelevant information depending on the query context.

3) user inexperience and inability concerning both the query subject as well as the search tool [8],
4) morphologic variants [9] - when relevant texts are not retrieved because the user did not use the morphologic variant (exed form or plural) used by repository texts.

Loh and Wives suggest solving these problems through recovery techniques using contexts or domains [10], where Context or Conceptual Space is understood as the set of words that characterize a subject or knowledge domain. Techniques based on document's terms frequency to determine the term relevance within the given document; how much is a term related with a context (how much it helps to define the context) and the relationship level between terms (to figure out the best terms to define a context) are subjects of research. These techniques help define which terms can be used to recover documents from a particular context.

A problem with such measures is that these techniques work better in large sized texts [11]. When analysing slides, text size must be taken into consideration because the text contained in them are of medium or small size. In this case, the size of the text is not proportional to the text's importance. Small texts can contain useful definitions, an observation about some subject, a solution to a problem and all these features are important to a learner. Furthermore, words do not repeat often in the text, invalidating the use of term frequency to determine relevance.

The IR process begins with an indexing task, where one or more descriptors are used to classify document content [8]. Most of these descriptors are nouns. Although natural language sentences are composed by nouns, pronouns, articles, verbs, adjectives and connectives, nouns are the grammar class that contains the highest semantic information while other grammar classes have phrasal complementary functions [11]. When noun groups are adopted as indexing terms, a conceptual logical view of the documents, composed of sets of non-elementary index terms, is obtained [5].

Terms that express content inside a text are called concepts. Indexing techniques are usually based on statistical measures that count the concepts existing in texts mapping them into an index structure called *inverted list* (also called reverse list or inverted index)[15]. To identify concepts, pre-processing techniques may be used such as eliminating *stopwords* - words with no meaning (articles, pronouns, etc.) or any word that does not belong to the class of nouns or cannot act as a noun in any kind of context/sentence – and stemming - that aims to identify the root concepts for grouping (Ex. Root of 'generation', 'generating' and 'generate' is 'generat').

Recovery is done consulting the inverted index to identify the texts containing the largest recovery coefficient.

# 3 Slide Information Retrieval System (SIRS)

The use of slides in the classroom allows teachers to organize and summarize the topics to be addressed, display graphics, images, videos, and other visual material to break the monotony of class.

The Slide Information Retrieval System (SIRS) allows students to retrieve relevant didactic material that can assist in understanding or complementing their knowledge regarding the subject being presented by the teacher in a slide presentation. In order to improve the tool's performance, the interpretation of the material displayed in the slides is of extreme importance.

Hurst says that information from the slides and annotation stream can be retrieved by just using the text from the slides to create a searchable index [2]. However, most traditional approaches for text indexing are based on the two assumptions that a term has a higher

relevance if it appears more often in a particular document than another term with less relevance (mathematically modelled through the Term Frequency, TF) and if it appears in rather few documents of the whole collection (mathematically modelled through the Inverse Document Frequency, IDF).

According to [2], such approaches that depend on the overall term distribution in a document collection can be critical if those documents contain mainly keywords instead of continuous sentences. In favour of a short and compact presentation, it is rather uncommon to use single terms more than once when preparing slides (even if they have a high relevance in a particular context). From this perspective, using an approach where more frequent words are extracted to represent the slide context is not efficient.

On the other hand, slides offer many additional information about the relevance of the content. Word that have been highlighted, are in italics or belong to the slide title usually represent important concepts.

The Slide Information Retrieval System (SIRS) is an add-in for Microsoft PowerPoint 2003/2007 that aims to further extend a slide presentation tool with contextualized document search capabilities, providing an electronic environment for classroom collaboration, besides being a tool to support the student in the process of self-learning, especially at times when additional literature is required.

Queries can be applied to a repository of documents previously selected by the teacher or to the Internet, using a browser. Priority is given to the repository as the texts have already been analysed and validated by the teacher. When the slide presentation is opened the content is automatically processed and the relevant terms extracted. These relevant terms are used to select the documents in the repository that are related to the content being presented. The relevant terms are also marked as hyperlinks in the slide texts. If the student clicks on the link, an expanded query is sent to a browser.

If the result is not satisfactory, WordNet [16] can be used for query expansion (also called Query Enrichment, Query Flexibilization, Query Contextualization). Query expansion is a technique used in information retrieval to better contextualize the original user query. Basically, the technique consists in the addition of related terms to the terms originally defined in the query. To enable query expansion, it is necessary to identify related terms, such as synonyms, stemming variations, or terms that are close to the query terms in the slide. Query expansion aims to define the query, disambiguating it, providing texts more consistent with the intention of the user when he types a query.

WordNet is a lexical database that organizes English nouns, verbs, adjectives, and adverbs into sets of synonyms, each representing a lexicalized concept. Semantic relations link the synonym sets [16].

Original query: **computer programming**

Query expansion: **computer programming program data language algorithm programmer** …

*Figure 1 - Example of query expansion.*

SIRS uses stopword lists and WordNet [16] to identify the nouns in the slides, to ensure the recovery of documents that are associated with the content being studied.

Stopword is any word that does not have enough semantic significance to define some domain or content. Are considered stopwords all the words belonging to the following grammar classes: prepositions, articles, determiners, etc. They are called by Miller closed-classes, while open-classes are nouns, verbs, adverbs and adjectives, that can function as nouns in a context, as adjectives in another context, etc. [16]

SIRS starts the concept extraction task by using a pre-defined stopword list. All words selected from the slide are sought in that list. If it matches, the word is not added to the index. If does not match, the word is submitted to the WordNet electronic dictionary to classify it as a concept or not, constructing the index. Following Baeza's recommendation, a word is considered a concept, in this project, when it can function as a noun in some given context [15].

After the indexing process is complete, the concepts stored in the index are used to expand queries and recover relevant material. Since concepts define the relevant content of the slide, this configures a contextualized search, because an implicit disambiguating of words is done when concepts related to a specific domain are used together in the query.

These expanded queries can then be applied to search a given repository or the Internet using Google for example. The results are displayed to the learners through links associated to the relevant concepts identified in the slides. Through a click, the users can easily have access to the recovered material.

Figure 2 shows the process in more details. In step 1, the process starts with the access to the slide, collecting all the words contained there and eliminating stopwords. Step 2 shows the words that were not eliminated and are therefore submitted to WordNet that extracts the nouns to identify concepts in step 3. Step 4 shows the concepts recovered from the slide, that are applied in a Google query, culminating in step 5, where material related to the initial slide content are displayed to the user (teacher or learner).

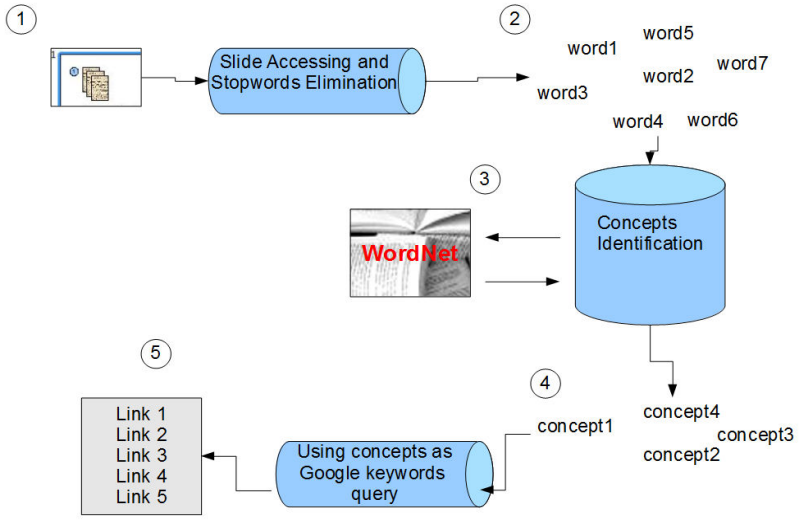



*Figure 2: The recovery process.*

# 4 Conclusion and Future Works

This paper presents the Slide Information Retrieval System (SIRS), a system that automatically contextualizes slide presentations to recover relevant material that can be used by students to complement their studies. The identification of the relevant concepts contained in the slides is done using stopword lists and WordNet, an electronic dictionary. The identified concepts are used to recover documents from a repository or from the Internet through the definition of expanded queries.

The use of this system allows the recovery of material that is always up-to-date. It removes from the teacher the need to constantly update links or reference lists associated to the subject presented in the slides. Every time the teacher finds material that he thinks is interesting he can include the material in the repository and the system will automatically associate it to the slide presentations related to the same context. As the association occurs when the slide presentation is opened, the recovered documents represent the latest version of the repository. Thus, if the student opens the presentation just after a relevant document is included, he will have access to it.

The identification of the relevant concepts helps to identify the domain in which the slide is inserted. Query contextualization helps to obtain results that are closer to users' aspirations. This significantly reduces the time spent by the student in the search for material that can be used for studying. This is important because with the ever increasing amount of material available in the Internet, the retrieval of relevant documents is very time consuming.

The implemented system confirms that contextualized searching is a good way to achieve better results than using just isolated keywords without any contextual treatment. And shows that nouns have, definitively, the ability to express content better than other grammar classes.

Even though the results are promising, improvements have already been envisioned. They include, for example, indexing looking at different colour or font style terms, as suggested by [2]. Other improvements include the selection of a smaller amount of nouns to define that domain. In this case, using adjectives can be a good way to improve results. Considering that adjectives qualify the nouns, they can properly define a domain. For example: "boolean logic", "electrical circuits", "organic chemistry", etc.

## References:


[1] Neder, M. L. C. and Possari, L. H. V. **Oficina para produção de material impresso**. Curso de formação em educação a distancia: Educação e comunicação em educação a distancia. 2001.

[2] Hurst, W.: **Indexing, Searching, and Skimming of Multimedia Documents Containing Recorded Lectures and Live Presentations**. Proceedings of the eleventh ACM international conference on Multimedia. Berkeley, USA. 2003.

[3] Anderson, R., Anderson, R., Davis, P., Linnell, N., Prince, C., Razmov, V., and Videon, F. **Classroom presenter: Enhancing interactive education with digital ink**. Pages 56–61. IEEE Computer.2007.

[4] Feitosa, A., **Organização da informação na Web Das Tags a Web Semântica**. Editora Thesaurus, 2006.

[5] Dridi, O., Ahmed, B. M.: **Building an Ontology-Based Framework For Semantic Information Retrieval: Application to Breast Cancer.** IEEEXplore. http://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=04530001. 2008.

[6] Brascher, M.: **A Ambiguidade na Recuperação da Informação**. DataGramaZero Revista de Ciência da Informação. 2002.

[7] Krovetz, R., Croft, W. B. **Lexical Ambiguity and Information Retrieval**. ACM Transactions on Information Systems, Vol. 10, No. 2. pp. 115 - 141. 1992.

[8] **Information about Google**. http://www.wbh.com.br/robogoogle.html. (2006)

[9] Cardoso, O. N. P.: **Recuperação de Informação**. http://www.dcc.ua.br/infocomp/artigos/v2.1/olinda.pdf , 2003.

[10] Loh, S., Wives, L. K., Frainer, A. S.: **Recuperação semântica de documentos textuais na internet**. CONFERENCIA LATINO-AMERICANA DE INFORMATICA (CLEI), XXV, 1999. Memorias del CLEI'99 (Tomo 2). Assuncion, Paraguay: Universidad Autonoma de Assuncion. 1999.

[11] Frakes, W. b., Baeza-Yates, R.: **Information Retrieval: Data Structured And Algorithms**. Prentice Hall PTR. 1992.

[12] Rosso P., Ferretti E., Jiménez D., Vidal V.: **Text Categorization And Information Retrieval Using WordNet Senses.** Proceedings of The Second Global WordNet Conference. Brno, Czech Republic. 2004.

[13] Gonzalo, J., Verdejo F, Chugur I, Cigarrán J. **Indexing with WordNet synsets can improve text retrieval.** Available in: <http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.14.4119>. Access date: Jun 20 2009.
[14] Silva, L. O., Ambrósio, A. P. L. **PowerInf – Collaborative Learning Environment**. WEI- CSBC 2009. Bento Gonçalves, Brazil.
[15] Baeza-Yates, R., Ribeiro-Neto, B. **Modern Information Retrieval**. Addison Wesley. ACM Press. 1999.
[16] Miller, G. A. **WordNet: A Lexical Database for English**. Communications of the ACM. 1995.

## Author(s):

Valdemar, Graciano Neto.
Federal University of Goiás, Informatics Institute.
Caixa Postal 131 – CEP 74001-970. Goiânia - GO – Brazil.
vgracianoneto@gmail.com

Luciana, Oliveira e Silva.
Federal University of Goiás, Informatics Institute.
Caixa Postal 131 – CEP 74001-970. Goiânia - GO – Brazil.
Luciana.go@gmail.com

Ana Paula, Ambrósio.
Federal University of Goiás, Informatics Institute.
Caixa Postal 131 – CEP 74001-970. Goiânia - GO – Brazil.
apaula@inf.ufg.br