MEETING LIFELONG EDUCATION REQUIREMENTS IN HANDLING UNSTRUCTURED INFORMATION THROUGH AN AUTOMATIC CONTENT HANDLING SYSTEM[1]

Athanasios G. Malamos

Department of Applied Informatics and Multimedia
Techonological Educational Institute of Crete
amalamos@epp.teicrete.gr

Georgios Mamakis

Department of Applied Informatics and Multimedia
Techonological Educational Institute of Crete
gmamakis@epp.teicrete.gr

Yannis Kaliakatsos

Department of Electronics
Techonological Educational Institute of Crete
giankal@admin.tecrete.gr

Anastasia Axaridou

Department of Applied Informatics and Multimedia
Techonological Educational Institute of Crete
aaxaridou@yahoo.gr

J Andrew Ware

School of Computing, University of Glamorgan
jaware@glam.ac.uk

Abstract

Education has always been a domain susceptible to changes initiated by technological advances. Still, modern educational schemes embed traditional approaches when it comes to lifelong education. The requirements imposed by the heterogeneity of the users in such cases are not taken under consideration. Thus, a whenever-wherever-whatever approach to accessing knowledge is needed to extend current educational models, so as to validate and promote lifelong education. However, lifelong education's requirements vary to such an extent that it is vital to be able find and handle information in a structured rather than in an unstructured form. Under this scope, we have developed an automatic summarization algorithm enabling access to unstructured knowledge by applying statistic and semantic techniques in order to fulfill lifelong education's requirements.

Keywords

Lifelong education, automatic content summarization.

[1] This work is co-funded by the Hellenic Ministry of Education and the European funds.

INTRODUCTION

Traditional educational schemes feature a teacher-class-student triplet, as the basis for accessing knowledge. The teacher is responsible for forming and supplying knowledge to the class which, apart from the set of students, is considered the virtual or real space where all the students along with the educator meet. As we have already mentioned in (Malamos et al., 2003) educational procedures, especially when referring to life-long education, need to be adjusted. The reasons for these changes derive from the categorization that can be applied to the learning audience regarding socioeconomic factors. An approach to verify the learners' categories, especially regarding lifelong education, can be considered to be:

- Employees and managers that need to rearrange their carrier
- Unemployed men and women
- Travelers
- Emigrants or people that belong to Ethnic Minorities
- People with disabilities
- Prisoners
- Older people

Judging from the aforementioned, we may stop referring to education as access to knowledge but we can consider learning procedure as a means of acquiring information. However, these categories present their own distinctive needs when regarding access to information. This problem tends to become more complicated when personal parameters are taken into consideration. Such personal needs are:

- Learning speed
- Available time
- Cost

Other personal parameters that affect modern educational procedures are:

- Family
- Occupation
- Educational background

The heterogeneity implied by the different number of possible user requirements, almost force the utilization of a wherever, whenever, whatever (WWW) approach (Malamos et al. 2005). A key factor that prevents educational schemes from achieving an ultimate encapsulation of lifelong education through a WWW approach is the fact that knowledge itself is unstructured. Also knowledge (Kalogianakis et al. 2005B) should be easily accessible. This is quite difficult when referring to unstructured information as, the user needs to identify between the information acquired and the one he really required. While in the case of a traditional class (Kalogianakis et al. 2005A), the educator is responsible for transforming and presenting the filtered knowledge to the learning audience, in terms of lifelong education, with all the parameters discussed, this can not be achieved. Structuring knowledge in such cases demands its prior classification, as lifelong learning audience mostly searches for specific pieces of knowledge and a step-by-step reformation of potentially accepted pieces of knowledge in order to ultimately become accepted by the learning audience, as unstructured knowledge is very difficult to handle. Also the system should be able to meet each respective user's demands and thus a personalized approach needs to be inherited with regards to these demands. Personalization in such terms can be achieved by the level of structured information presented to the user each time. Under this scope, we have developed an automatic summarization algorithm based on statistic and semantic techniques that classifies according to subject, and generates an extract of any given document according to that subject. The main ideas behind the development of the algorithm derive from the assumption proposed by Edmundson in (Edmundson, 1969), along with research introduced ever since, utilizing semantics in the area of text manipulation by means of statistical or syntactical analysis [(Baxendale,1958), (Saggion, 1999), (Teufel & Moens, 1999),(Ono et al., 1994) and (Barzilay & Elhaddad, 1997)].

MOTIVATION

As implied by the aforementioned, modern e-learning applications should evolve to meet lifelong learning requirements, as acquired by the complexities imposed by the underlying heterogeneity. First of all, the system should structure knowledge according to the user requirements, since unstructured knowledge is difficult to handle. It should also provide quick access to the desired results independently of the kind of information requested. Thus, it should be adaptable to fully encapsulate and support dynamically updated subjects (or thematic domains), through the utilization of respective mechanisms. This, as opposed to traditional content handling systems based on newspaper corpora of documents, requires an adaptive and automatic keyword extraction mechanism that would not only identify static keywords already classified in the system's operational scope, but also encapsulate potentially unclassified keywords to form new domains targeting on education. This stems from the evolving nature of knowledge itself, especially in cases as sciences, where new scientific results are provided daily. Our motivation is to provide an application offering personalized services, in order to satisfy each potential user, with the primary target of providing knowledge in a user-customizable, structured way.

METHODOLOGY

In order to develop a platform for automatic content summarization for lifelong education, it is vital to identify and validate the system requirements, as imposed both by the desired form of representation and by the potential users. This led us to the extension of a, currently under development, automatic content summarization system proposed by the authors (Mamakis et al., 2005) to include limitations and requirements imposed by lifelong education. The system is based on a combination of four inter-connected algorithms, widely used in the area of content management, with the use of statistical information acquired from corpora data: an extended stemming algorithm capable of extracting roots of words while simultaneously identifying nouns present in each grammatical form in a text (as they are considered vital for our document summarization system (Mamakis et al., 2005)), a thematic domain algorithm capable of creating sets of words that share higher proportional percentage in coexisting in a document, a domain classification algorithm capable of categorizing an article into one or more of the thematic areas identifying by the system, and an extractive summarization module, capable of extracting in a step-by-step manner a summary of a classified document.

This system was enhanced with features in order to fulfill the requirements imposed by lifelong education. Thus, it is able to identify a large variety of topics of potential interest. The resulting extract's length can be user-defined, implying a customizable level of detail.

ALGORITHMS CREATED

Stemming module

As stated before, we have identified that nouns are important word elements in the extraction of a summary. The stemming algorithm we have developed is targeted on extracting stems of nouns regardless of their clause and disregarding other word elements, and is based on portions of Porter's stemming algorithm (Porter, 1980) and the Greek stemming algorithm proposed by Kalamboukis in (Kalamboukis, 1995). Greek noun endings are shown in Table 1. Still noun endings are not enough in extracting a noun as active voice participles, passive voice participles and adjectives may interfere with the isolation of a noun. Our stemming algorithm can discriminate between nouns and active voice participles and passive voice participles, but fails when referring to adjectives. Still since adjectives are not domain specific words, an efficient corpus of documents for the creation of the domains will disregard them.

Table 1. Accumulative table of Greek noun endings in all forms.

Noun genders	Possible Noun Endings
Masculine	ας, α, αδες, αδων, ες, ων, ης, η, ηδες, ηδων, ες, εδες, εδων, ους, ουδες, ουδων, ος, ου, ο, ε, οι
Feminine	α, ας, ων, ες, αδες, αδων, η, ης, ες, εις, εων, ω, ως, ος, ου, ο, οι, ων, ους
Neutral	ο, ου, α, ων, ατα, ατων, ι, ιου, ια, ιων, υ, ιου, ος, ους, η, , –α, ατος, ας, ως, ωτος, των

The same algorithm is used both to create the semantic domains and to isolate the nouns of a document.

Off-line Step – Domain Creation

Our domain creation algorithm takes advantage of the occurrence of a noun in a document of the corpus. If a noun appears in all domains then it is disregarded as insignificant.

Document Classification algorithm

In order to effectively extract a summary of an article, it is vital to identify the thematic subject of the document, in order to verify the semantics of the words used and calibrate their use. We adopt a statistic approach to verify the domain the document at hand. We calculate a similarity factor, according to formula:

$$sf = \sum \frac{\sum_{i=0} \frac{A_i}{L_i}}{N_j}$$

where A the number of occurrences of word in the document, L the total nouns of the i-th document, and N the total number of nouns in the j-th domain. Thus, the domain the document belongs to is the one acquiring the maximum sf. The initial step used in this algorithm is, as in domain creation step, our stemming algorithm. The resulting noun stems constitute the set of words compared to the noun stems of the domain.

Step 3 – Document extraction

The final step of the algorithm, extracts the summary of the document, in a statistic and domain-oriented approach. Factor hr is computed as:

$$hr= f \cap D /f$$

where f represents the set of noun stems of a sentence of the document and D the set of noun stems constituting the domain. hr factor along with the absolute position of a sentence in the document constitute a hash table. The summary is extracted by constructing a document with the sentences having the greater hr factors in their relative position to one another. The number of sentences to be included is decided by the number of words, the user wants to include in the resulting summary.

CONCLUSION-FUTURE WORK

In our future plans, we study the utilization of the algorithm in a knowledge based portal, where, instead of offering links to web pages that are potentially of interest, as in conventional web-based search engines, we would further extend keyword-based search with summarized results, according to the end-users' specifications. In order for that to be accomplished, we consider further refinements primarily to the stemming and domain creation algorithm as it was observed that certain domains may be formed by a very large set of words, thus becoming generalized. Also, we consider applying an evaluation system to measure the efficiency of our algorithm as opposed to other summary extraction algorithms for different languages.

REFERENCES

Barzilay, R, and Elhadad M. 1997. "*Using Lexical Chains for Text Summarization*". In Proceedings of the ACL/EACL'97 Workshop on Intelligent Scalable Text Summarization, 10–17. Madrid, Spain

Baxendale, P.B. 1958. "*Machine-made Index for Technical Litterature - an experiment*". IBM J. Res. Dev. 2(4): 354–361

Edmundson, H.P. 1969. "*New Methods in Automatic Extracting*". Journal of the Association for Computing Machinery 16(2): 264–285

Kalamboukis T.Z.(1995), "*Suffix stripping with Modern Greek*", Program 29(3), pp 313-321

Kalogiannakis M, Psarros M, Liodakis G, Vassilakis K. "*Asynchronous tele-education:*

main or complementary tool of the course? First perceptions of students and teachers at Technological Education Institute (TEI) of Crete", , Proceedings of the Annual Conference on Telecommunications & Multimedia, 311-317, TEMU 2005. (2005 A)

Kalogiannakis M, Vassilakis K, Psaros M, Lionarakis, A. "*ICT and Pedagogical framework in Distant Learning*", in Proceedings of the 3rd International Conference on Open and Distance Learning, Vol. A', 481-496, ICODL 2005 (2005 B)

Malamos A.G., Kaliakatsos Y., Axaridou A., "*Improving training by applying WWW (Whenever- Wherever- Whatever) capabilities to learning platforms.*" Proceedings of 3rd International Conference on New Horizons in Industry and Education, Santorini, 28-29 August 2003

Mamakis G., Malamos A.G., Kaliakatsos Y., Axaridou A., Ware J.A. "*An Automatic Content Summarization algorithm for Greek Language*", presented in ICICT 05, 5-6 December 2005 Cairo, Egypt

Ono, K., K. Sumita, and S. Miike. 1994. "*Abstract Generation Based on Rhetorical Structure Extraction*". In Proceedings of the International Conference on Computational Linguistics, 344–348. Kyoto, Japan.

Porter, M.F., 1980, "*An algorithm for suffix stripping*", Program, 14(3) :130-137

Saggion, Horacio. 1999. "*Using Linguistic Knowledge in Automatic Abstracting*". In Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics, 596–601. Maryland, USA

Teufel, S, and Moens M. 1999. "*Argumentative classification of extracted sentences as a first step towards flexible abstracting*". In I. Mani and M.T. Maybury, eds., Advances in Automatic Text Summarization, 155–171. The MIT Press.