A Framework for Content Preparation to Support Open-Corpus Adaptive Hypermedia

Killian Levacher, Éamonn Hynes, Séamus Lawless, Alexander O'Connor, Vincent Wade

Centre for Next Generation Localisation, Knowledge & Data Engineering Group,
School of Computer Science and Statistics, Trinity College, Dublin, Ireland

{Killian.Levacher, hynese, Seamus.Lawless, Alex.OConnor,
Vincent.Wade}@cs.tcd.ie

Abstract. A key impediment for enabling the mainstream adoption of Adaptive Hypermedia for web applications and corporate websites is the difficulty in repurposing existing content for such delivery systems. This paper proposes a novel framework for open-corpus content preparation, making it usable for adaptive hypermedia systems. The proposed framework processes documents drawn from both open (i.e. web) and closed corpora, producing coherent conceptual sections of text with associated descriptive metadata. The solution bridges the gap between information resources and information requirements of adaptive systems by adopting state-of-the-art information extraction and structural content analysis techniques. The result is an on-demand provision of tailored, atomic information objects called "slices". The challenges associated with open corpus content reusability are addressed with the aim of improving the scalability and interoperability of adaptive systems. This paper proposes an initial architecture for such a framework in addition to reviews of associated technologies.

Key words: Open Corpus Content, Adaptive Hypermedia, Statistical Content Analysis, Structural Content Analysis

1 Introduction

The increasing pervasiveness of the internet is fundamentally changing how people author, interact with and consume content. There are early signs of a shift in the way digital content is created, from the linear authoring and publication of material towards the aggregation and re-use of existing content from various disparate sources.

Adaptive Hypermedia Systems (AHS) have traditionally attempted to deliver dynamically adapted and personalised presentations to users through the sequencing of reconfigurable pieces of information.

While the effectiveness and benefits of such systems have been proven in numerous studies [1][2], a major obstacle to their widespread adoption relates to the restriction of suitable content available to provide such adaptivity in terms of volume, granularity, style and meta-data.

One of the key impediments in offering a wide range of information objects is, for example, the considerable amount of manual effort involved in creating these resources. Information object content consists of either pre-existing documents [3], or of content created by small groups of individuals [4], intentionally authored for a particular system [5]. This generates both scalability and interoperability issues, which prevents the mainstream adoption of adaptive systems.

Metadata standards such as LOM [6] attempt to provide generic content packaging methods with the aim of enabling the interoperability of information objects within AHSs. Although they implement usage-agnostic principles, their development is very time consuming due to the complexity involved [7]. Furthermore, they also require AHSs to comply with a specific content structure in order to make full usage of these resources.

In parallel with these developments, a wealth of information is now accessible on the web, in digital repositories and as part of library initiatives. However, due to its heterogeneity, this information is not suited to direct usage by AHSs in its present form. It is available in several languages and within large documents with very limited meta-data. These objects are also generally very coarse-grained and correlated with unnecessary noise such as navigation bars, advertisements etc. Automated content preparation techniques are therefore necessary to provide scalable solutions to AHS specific content requirements.

Our goal is to provide AHSs with the ability to benefit from the wealth of knowledge already accessible on the web and in digital libraries by bridging the gap between these information resources and the information requirements of adaptive systems. We propose the creation of a service that can tailor open-corpus content for use in AHSs without requiring such systems to adopt a generic content structure. This system would hence resolve scalability issues in content authoring by providing AHSs with a large array of information objects. Moreover, the inclusion of open corpus information would provide up-to-date content in a wide variety of styles and structures.

We aim to combine several disciplines such as Information Retrieval, Adaptive Web and Information Extraction in order to provide an on-demand information object service. This service will harvest open corpus information, segment it structurally and automatically augment it with successive layers of internal metadata. Any AHS could then provide this service with a request that includes requirements for an information object which would in turn fetch and tailor this content to its specification.

This paper is structured as follows: Section 2 will present the key challenges addressed within this research in relation to AHS information requirements as well as structural and statistical text analysis. The framework we envisage developing is presented in detail in section 3, along with the overall workflow produced. Section 4 will present the application area of such a technology and how we intend to evaluate it. Finally, section 5 will conclude and present the road map ahead. This work is applicable to any type of content, however some of the examples presented in this paper will be based on adaptive e-learning systems.

2 Key Challenges and Related Work

Throughout the evolution of adaptive systems, technologies introduced have moved from original "all in one" solutions, towards increasing modularity whereby AHSs become consumers of domain, pedagogical and user models [8]. This has the effect of leaving such systems with a core adaptive engine capable of dealing with a wider range of loosely coupled models that are integrated as desired. Content, however, is still very tightly coupled to these engines and as a result strongly impedes the general adoption of these systems.

Most AH systems, up until recently have operated in closed document spaces with content specifically authored for their usage [4], hence obstructing interoperability, both by accepting only a narrow field of content, and by motivating the generation of content in a highly-specific format. As a result adaptive systems also encounter scalability issues due to the limited amount of content available to adapt on, arising from the small amount of manual contributions available. Open corpus content is increasingly seen as providing a solution to these issues [9]. However, most systems incorporating this type of content have, for the moment, mainly focused on linking it with internal content as alternative exploration paths. Those incorporating it fully into their system [10][11] require manual mark-up of such content with specific meta-data. Schemas such as LOM (in the area of e-learning) and IMS packagings, attempt to provide usage-agnostic solutions, however they require a lot of development effort thus prohibiting scalability [7]. Moreover, AHSs must comply with specific content structures so as to avail of these resources. In order to leverage the full potential of open corpus resources, adaptive systems need to incorporate and correlate this type of content fully into existing systems using structure-agnostic and automated approaches.

Our approach, on the other hand, moves away from the packaging model altogether by considering content as a pluggable feature of adaptive systems to the same extent as user or domain models. A service providing open corpus content automatically in a form and format customised for each requesting adaptive system decouple this feature from current systems, disentangling it from content provision and hence providing a solution to both scalability and interoperability issues.

Fully integrating information harvested over the web within adaptive presentations, on the other hand, generates new issues. Web pages are usually written as stand-alone content without any re-composition purposes. In addition to the main content, they usually present navigation bars, advertisement banners and irrelevant features from various sources that must be removed prior to proper re-use.

Structural segmentation will therefore be an initial fundamental requirement in tailoring open corpus content for use within adaptive systems. A large proportion of pages harvested will need to be stripped of any redundant information that might prevent its adequate re-use in adaptive systems.

The problem of structural segmentation has already been addressed from different perspectives mainly with the purpose of tailoring existing content for different displays [12]. Most of the techniques up until recently have focused on analyzing the DOM structure of a HTML page as a basis for segmentation. Several attempts, for instance, focus on removing templates from pages by identifying common DOM sub-trees [13] or using isotonic regression [14]. Machine Learning algorithms are also used to decide which pair of tags should coincide with suitable segmentation points. However, it has become increasingly popular to separate any formatting style from within HTML tags using Cascading Style Sheets (CSS) and Javascript, thus increasing the heterogeneity of rendered content from similar HTML trees [15]. For this reason, such techniques do not appear to be an adequate solution when dealing with a large set of documents as diverse as the World Wide Web.

Vision-based techniques, using entropy reduction [12], or techniques such as VIP algorithms [16] on the other hand, partition a page based on its rendering. These have the benefit of covering a wider range of pages regardless of their HTML trees. But their usage within systems dealing with large volume of data is questionable since rendering must be performed prior to any analysis, inevitably producing delays. Moreover, they completely ignore DOM features altogether which do provide structural clues where rendered surfaces appear similar [12].

This part of our research will therefore seek to provide a solution to this paradigm that scales both in terms of processing speed as well as structural segmentation precision within a wide heterogeneous range of pages. The procedure will additionally need to incorporate the notion of granularity in order to keep cohesive chunks of text together to provide meaningful input to subsequent semantic analysis.

Within the statistical analysis component of the framework, it is important to extract concepts with high precision/recall and efficiently represent these concepts in such a way as to make them as interoperable and re-usable as possible. Brusilovosky *et al.* [9] discuss the lack of re-usability and interoperability between adaptive systems and how current adaptive hypermedia systems are restricted to closed corpora by the nature of their design. This framework architecture is based on the idea that open corpus adaptive hypermedia is feasible: the effort will be in moving away from tables of document-term frequencies towards a richer semantic representation of data that is uncomplicated, easy-to-use and highly interoperable.

3 Framework for Content Preparation

The framework proposed in this research is divided into three separate components as shown in figure 1. Each part of the framework executes a specific task on the open corpus content. A specific layer of meta-data is subsequently appended, enriching it with structural and semantic clues for further analysis at each step. The first two stages of the pipeline are executed pre-emptively prior to any client request while the intelligent slicing task is executed at run time.

The concept of a "slice" is an abstract notion representing a stand-alone piece of information, originally part of an existing document, extracted and segmented to fulfil a specific information request. A slice can be atomic or composed of other slices. It possesses its own set of properties (or meta-data) and can inherit those of others. A slice is virtual in the sense that it only exists temporarily to fulfil a client request. It is specific to each information request and represents a subjective perspective on a particular piece of a document and its description. The degree of complexity of a slice will match the requirements of the system which requests it.

The rest of this section will present in detail each stage of the framework. The structural analysis of open corpus content will initially be described followed by the statistical analysis of these resources. Finally, the overall work-flow between this framework and AHSs clients will be outlined using intelligent slicing technology.

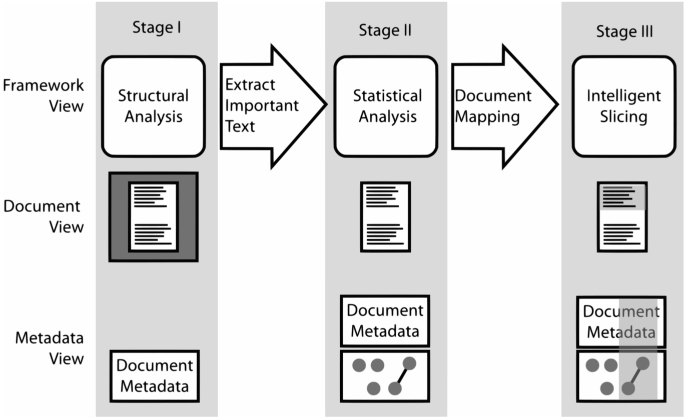

Fig. 1. Content analysis methodology

3.1 Structural Analysis

In order to provide the initial structural segmentation needed to create slices, this research will attempt to combine benefits of both DOM- and vision-based segmentation techniques. The process will initially divide pages in atomic blocks using relevant HTML tags similar to [15]. Following this extensive break down, the page will be reconstructed, aggregating relevant blocks together with similar inherent textual properties. Changes in text flow (short phrases in navigation bars vs. long sentences in text), for instance, will be considered as segment delineations.

The notion of text density, originating from the field of computer vision will also be applied to these pages using the ratio between the number of words and the space occupied on the page [15] as an additional hint for section divisions. Furthermore, the use of linguistic analysis techniques such as style boundary detection will be considered as additional boundary indication.

3.2 Statistical Analysis

In this section we discuss methods for statistically analysing text and extracting information such that it can be used as part of the wider framework; it serving as a substrate for the incorporation of several different statistical analysis techniques. In order for the proposed framework to be able to interact naturally with information seekers, it must cope with inherent ambivalence, impreciseness or even vagueness of requests and deal with them in a structured and systematic fashion. For these reasons, statistical methods offer a wealth of neat techniques for dealing with the underlying complex characteristics of language.

Work currently underway includes looking at methods for the automatic extraction of concepts from text. The vision is to extract and create a conceptual layer over the text which is amenable to efficient searching and logical reasoning. We move away from working only on closed-corpus texts and go some way to addressing the huge challenge that is the open domain problem [9] by making our algorithms as generalised as possible and not restricted to any particular domain. One method for identifying concepts from text is to use supervised learning methods. We use HMMs (Hidden Markov Models) to extract coherent, relevant passages from text [17]. This method is particularly suited to extraction from expositional texts as opposed to highly expressive literary texts (i.e. poetry, literary prose, etc.) which have high entropy vocabularies and diverse language use. Jiang *et al.* report very high precision/recall compared to other methods and their work fits in nicely with the idea of extracting concepts/coherent relevant passages of text.

The basic idea is as follows: given some *a priori* concept that is to be extracted from a text, several passages identified as being associated with that particular concept are used as training data for the HMM classifier. Once a reasonable amount of training data has been collected (enough data such that the classifier's train and test error converges towards the desired performance), the HMM can then make soft decisions on a text stream and identify coherent relevant passages of varying length with very high accuracy.

The topology of the HMM is shown in figure 2 below: the states B1, B2 and B3 are the "background" states, state E is an end state and state R is a "relevant" state. The 5 state HMM structure allows the system to emit a relevant symbol, move to a background state and then re-emit a relevant symbol at some later point before reaching the E state. The output observation probabilities can be extracted directly from the training corpora, and the objective is to learn the transition probabilities so as to maximise the likelihood of observing a relevant passage given a concept.

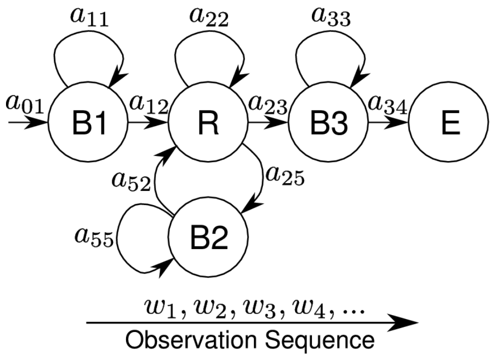

Fig. 2. HMM topology

On a synthesised corpus consisting of a number of article abstracts (coherent relevant passages) stitched together, with each abstract having an associated label, very high precision and recall values were reported [17]. It remains to be seen how well the HMM method performs on non-synthetic corpora and work is currently underway to explore this.

Several extensions to the work of Jiang *et al.* have also been explored, including the incorporation of anaphora to resolve pronouns and augment the presence of important entities such as people and places. In addition to the uni-gram lexical features, more elaborate query expansion methods and semantic analysis has been looked at to extract more elaborate features for the classifier to work from.

The incorporation of this additional functionality is expected to yield higher performance in terms of precision and recall on the synthetic document set and more reliable performance on real corpora. Instead of just a single-dimension vector with just one feature (i.e. a stemmed word), the word is represented as a slightly higher dimensional vector (i.e. the stemmed word with additional features) that helps improve the overall accuracy of the classifier.

Word Sense Clustering Another method used to extract conceptual entities from text is word sense clustering. Lexical resources such as WordNet [18] provide fine-grained sense descriptions for a huge number of words in the English language. However, such fine-grained detail is not always required: it is possible to cluster together word senses and come up with a looser word sense that abstracts away a lot of the detail and is more compatible with the framework as described in this paper. Snow *et al.* [19] describe a supervised learning method for merging word senses using a SVM-based (Support Vector Machine) clustering algorithm. This algorithm yields senses at a slightly higher level of granularity and is more compatible with corporation into a conceptual layer representation. Use of latent semantic indexing techniques is another powerful and elegant method for dealing with synonymy and polysemy whilst at the same time, automatically generating concepts, or clusters of word senses [20]. In latent semantic indexing, each document is represented as a vector of terms, with each element of the vector having an associated weight; some function of the term frequency. Dimensionality reduction techniques are then applied to each of these document vectors, resulting in a set of vectors that point in the direction of semantically similar words. Each term in a document can then be projected onto the reduced-dimensionality semantic space, where the dominant vector(s) correspond to the concept that the term is most associated with. Probabilistic latent semantic indexing techniques [21] provide an even more accurate means of mapping from terms to concepts and such techniques will be used extensively for the statistical analysis component of the framework.

Representing content in the form of a semantic map is thought to mimic the way the human mind classifies linguistic categories. While there is no direct link between semantic maps and the representation of concepts in the human mind, it is an objective starting point for investigation of the efficiency of the semantic map representation on human cognition.

3.3 Content Preparation Work-Flow

The framework architecture presented in figure 3 belongs to the core functional unit of a web service that delivers on-demand slices to adaptive systems.

As a content provider, the framework will first of all require domain specific data gathered previously by pre-emptive crawls in areas of interest to adaptive engine consumers. We envisage as an initial step to use the OCCS harvesting system [22], developed previously, as a content gathering tool in order to bootstrap this framework with a renewable wide range of content. The system will be content harvester agnostic, hence any other content gathering system could be plugged in if required, such as private repositories for instance.

Once suitable content is available, the framework described previously will firstly remove any redundant content, then hand over only relevant information to our semantic analysis unit, and then create a conceptual map of the data. These two steps will produce a large amount of internal meta-data, both from a semantic and structural point of view.

At this stage, the service will be ready to receive content requests from a diverse range of consumers through conceptual queries which the service will attempt to map as closely as possible to the internal meta-data gathered previously on its content. The intelligent slicing unit will be responsible for matching requests to available data within the system and slicing information objects to the finest semantic grain possible matching the query. Semantic maps constructed *a priori* will be used to produce a personalised slice for the adaptive systems. Semantic technologies (OWL and SPARQL) will form the underlying mechanism navigating through pre-emptive semantic mappings.

4 Evaluation and Roadmap

This section provides a brief summary of what are the main benefits of this framework in comparison to traditional systems. The limitations of this architecture are also discussed along with possible solutions. This will result in a

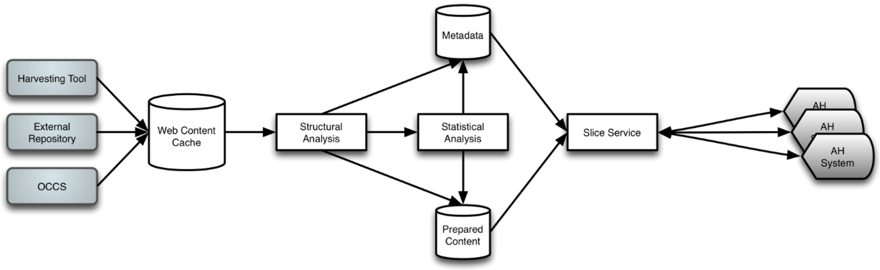

Fig. 3. Content analysis architecture

roadmap which focuses on an initial empirical evaluation of critical components, its implementation and the overall evaluation of the framework.

The content preparation framework discussed in this paper aims at bridging the gap between an increasingly large amount of open-corpus resources available on the web and AHS information requirements. The automated content slicing approach proposed offers a solution to the authorship scalability issues currently impeding the mainstream adoption of adaptive web technologies.

This method removes any content packaging interoperability issues by localising slices to the requirements of the AHS, instead of the AHS complying to a specific format. Moreover, the pipelined architecture of this framework enables the incorporation of additional plug-able meta-data generation services to the system, thus removing the need for unscalable hand-made meta-data.

However, this novel content preparation solution raises new challenges. At a component level, several technological challenges exist. To the best of our knowledge, no existing structural segmentation approach currently fulfills all of our requirements. For this reason, an initial evaluation of existing algorithms will be conducted, resulting in the combination of several techniques. Additionally, the semantic analysing step of this framework seen as the critical stage of this pipeline. An empirical evaluation of this step will therefore be needed so as to measure the quality and size of the meta-data created by this component. This process is critical since it will subsequently determine methods available to match slices with content requests. As a result, the content precision of slices generated will depend on the successive layers of the automated meta-data available. It is possible, in addition, that this precision could be improved by the use of an *a posteriori* manual crowd sourcing refinement of the slices.

As part of the Centre for Next Generation Localisation (CNGL), an initial implementation of this framework within a personalised multi-lingual customer care (PMCC) system is planned. A number of framework level evaluations will be undertaken in partnership with our industrial partners. An empirical evaluation of automatic meta-data generation performance will be conducted in order to estimate the degree of reliance towards automatic slice generation. These initial steps will be necessary to ultimately provide a reliable foundation for the PMCC application envisaged.

5 Discussion

This paper described how restrictions in the provision of suitable content for Adaptive Hypermedia Systems currently impedes their full mainstream adoption due to interoperability and scalability authoring issues. The decoupling of content infrastructures from Adaptive Systems along with the use of automatically tailored open corpus resources was presented as a solution to this predicament.

Consequently, the architecture for a content provision service that fills the gap between open corpus information resources and Adaptive System information requirements, was outlined. To the best of our knowledge, this framework represents the first attempt to produce a pipelined, client agnostic, information preparation service that produces virtual content slices for adaptive systems. Other innovations in the framework architecture include the modular and pluggable open corpus analysers, which produce successive layers of internal metadata information. The combination of these components and how they integrate coherently to produce the framework is also an aspect under investigation.

An initial prototype of this framework is currently under development and aims towards integrating fully with a customer care information system that is currently being deployed as part of the Centre for Next Generation Localisation project.

Acknowledgements This research is supported by the Science Foundation Ireland (grant 07/CE/I1142) as part of the Centre for Next Generation Localisation (www.cngl.ie) at Trinity College, Dublin. The authors wish to thank the anonymous reviewers for their comments.

References

1. Brusilovsky, P.: Adaptive Navigation Support in Educational Hypermedia: An Evaluation of the ISIS-Tutor. Journal of Computing and Information Technology (1998)
2. Hook, K.: Evaluating the utility and usability of an adaptive hypermedia system. International Conference on Intelligent User Interfaces (1997)
3. Henze, N., Nejdl, W.: Adaptivity in the KBS Hyperbook System. 2nd Workshop on Adaptive Systems and User Modeling on the WWW (1999)
4. Dieberger, A., Guzdial, M.: Coweb - Experiences with Collaborative Web Spaces. Interacting with Social Information Spaces (2002)
5. De Bra, P.: Teaching Hypertext and Hypermedia through the Web. Journal of Universal Computer Science **2** (1996)
6. IMS Global Learning Consortium: Learning Object Metadata (LOM) Information Model 1.2.2 (2009)

7. Farrell, R.G., Liburd, S.D., Thomas, J.C.: Dynamic Assembly of Learning Objects. In: Proceedings of the 13th International World Wide Web Conference on Alternate Track Papers & Posters. (2004) 162–169
8. Conlan, O., Wade, V., Bruen, C., Gargan, M.: Multi-model, Metadata Driven Approach to Adaptive Hypermedia Services for Personalized eLearning. In: Adaptive Hypermedia and Adaptive Web-Based Systems. Springer-Verlag (2002) 100–111
9. Brusilovsky, P., Henze, N.: The Adaptive Web. In: Open Corpus Adaptive Educational Hypermedia. Springer (2007) 671–696
10. Henze, N., Lenski, W., Wette-Roch, E.: Adaptation in Open Corpus Hypermedia. Journal of Artificial Intelligence in Education **12** (2001) 325–350
11. Carmona, C., Bueno, D., Guzman, E., Conejo, R.: SIGUE: Making Web Courses Adaptive. In: Adaptive Hypermedia and Adaptive Web-Based Systems. (2002)
12. Baluja, S.: Browsing on Small Screens: Recasting Web-Page Segmentation into an Efficient Machine Learning Framework. In: Proceedings of WWW-2006, ACM (2006) 33–42
13. Vieira, K., da Silva, A.S., Pinto, N., de Moura, E.S., Cavalcanti, J.M., Freire, J.: A Fast and Robust Method for Web Page Template Detection and Removal. In: Proceedings of the 15th ACM International Conference on Information and Knowledge Management, ACM (2006) 258–267
14. Chakrabarti, D., Kumar, R., Punera, K.: Page-Level Template Detection via Isotonic Smoothing. In: Proceedings of the 16th International Conference on World Wide Web, ACM (2007) 61–70
15. Kohlsch, C., Nejdl, W.: A Densitometric Approach to Web Page Segmentation. In: Proceedings of the 17th ACM Conference on Information and Knowledge Management, ACM (2008) 1173–1182
16. Cai, D., Yu, S., Wen, J.R., Ma, W.Y.: Extracting Content Structure for Web Pages Based on Visual Representation. Springer (2003)
17. Jiang, J., Zhai, C.: Extraction of Coherent Relevant Passages using Hidden Markov Models. ACM Transactions on Information Systems (TOIS) **24** (2006) 295–319
18. Fellbaum, C.: WordNet: An Electronic Lexical Database. MIT Press (1998)
19. Snow, R., Prakash, S., Jurafsky, D., Ng, A.Y.: Learning to Merge Word Senses. In: EMNLP. (2007)
20. Deerwester, S., Dumais, S., Furnas, G., Landauer, T., Harshaman, R.: Indexing by Latent Semantic Analysis. Journal of the American Society for Information Science **41** (1990) 391–407
21. Hofmann, T.: Probabilistic Latent Semantic Indexing. In: 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. (1999)
22. Lawless, S., Hederman, L., Wade, V.: Occs: Enabling the Dynamic Discovery, Harvesting and Delivery of Educational Content from Open Corpus Sources. In: Eighth IEEE International Conference on Advanced Learning Technologies, IEEE Computer Society (2008) 676–678