Integrated Copy-Paste Checking: Design and Services

Narayanan Kulathuramaiyer¹, Bilal Zaka², Denis Helic³
Faculty of Computer Science and Information Technology, Universiti Malaysia Sarawak, Malaysia¹
Institute for Information Systems and Computer Media, Graz University of Technology, Austria¹,²,³
{nara¹, bzaka², dhelic³} @iicm.edu

Abstract

The advances in technology have made academic cheating far too easy for learners. Furthermore, the World-Wide-Web has brought about a widespread culture of easy-access to all sorts of information, thus reducing the need for learners to perform diligent research or study. E-learning systems would then need to incorporate the monitoring and checking for student expressions of reading and writing, while guiding them towards learning the rightful skills. This paper describes the architecture and design of an integrated Copy-Paste system aimed at addressing these concerns.

1. Introduction

Advances in technology, particularly the Web that has made it far too easy for learners to commit unethical practices. The over-reliance of students on Web resources such as Wikipedia and Google without considering the reliability has become a major concern. In current teaching-learning systems, the identification of problem situations in student learning processes, are largely performed after they actually happen. This approach is however ineffective in addressing the needs of future learning environments.

A learning ecosystem called ICARE [1] has been proposed to overcome the problems mentioned and to serve as a model for future e-learning systems. The term ecosystem as used here refers to a community of learners and instructors, interacting cooperatively and collaboratively supported by a technology enabled learning environment. Details of the ICARE concept can be found in [2]. Key ideas of this ecosystem are summarized as follows:

- Support for students to be actively engaged in creating their own intellectual property and respect other people’s copyright.
- Provide a guided environment to foster constructive learning practices and critical thinking whereby students can structure their reading and construct written works systematically.
- Support for the acquisition of basic reading-writing skills i.e. paraphrasing, summarizing and referencing accurately.
- Support for process management to structure course-work as a series of process steps. Student learning can then be continuously checked and assessed at each stage.
- Incorporate preventive measures for making sure that unwanted versions of copy-and-paste just cannot happen (or is drastically minimized)
- Incorporate viable technologies required to minimize the supervision effort required from instructors.

This paper then further describes the architecture and design of an integrated Copy-Paste system, which is a key innovation of ICARE.

2. Overview of the ICARE Ecosystem

The ICARE ecosystem is being built on top of WBT-Master [3] [4], a sophisticated e-learning system that supports the definition of multiple learning scenarios, project-based administration of e-learning and the interactive classroom management activities such as mentoring, brainstorming, project management etc. [4]. An overview of the ICARE ecosystem is shown in Figure 1. It adopts a learner-centered approach in supporting learners’ acquisition of knowledge and skills. Multiple learning scenarios are used to implement alternative learning modes made available to students.
A rigorous academic reading and writing process is thus enforced in a guided environment. Each student maintains a portfolio, which characterizes and represents all learning outcomes, recognized student works and achievements. An e-diary is employed to enable the aggregation of student (individual and collaborative) contributions to be captured, assessed and reflected upon. Internal processes and states of the learners can then be represented and augmented with systemic input to provide deeper insights on their learning.

The ecosystem allows the demonstration of student learning and understanding via knowledge maps (K-Maps). K-Maps can be associated with any particular stage in the learning process, enabling instructors to determine the actual learning of students to make an informed judgment on possible cheating on the part of students. This ecosystem also allows the instructor or mentor to identify areas in which students found difficulty in understanding or assimilation. The Copy-Paste checking subsystem has been integrated directly into the e-learning environment in such a way that it directly monitors student activities and provides appropriate feedback to instructors, mentors and students themselves. We will describe the overall system architecture of ICARE before focusing on the Copy-Paste checking facility.

3. Architecture of ICARE

A mashup [5] is an emerging application development paradigm on the Web. It enables the rapid development of flexible applications built upon a collection of Web Services. A mashup architecture has been proposed for ICARE, built upon a set of Web services as shown in Figure 2.

There are a number of services, which are required to facilitate learning management in ICARE. These services can be separated into basic services such as K-Card management, profile management, sessions management, etc. and application specific services which includes concept discovery management, expertise finding, knowledge visualization, Copy-Paste administration, etc. Current works at Graz University of Technology include the development of a number of these Web services, as described in [6], [7], [8], [9].
has been pushed to the user in a non-obtrusive manner, the system continuously analyses the feedback of users (students, peers, mentors, etc.) to determine the value of the information supply. Based on a deeper analysis, students are advised and supported appropriately. The results of the analysis are then visualized and presented to the student.

Figure 3: Layered Web services Architecture

As opposed to current plagiarism detection methods as used in educational institutions, [10], [11], a conceptual similarity checking approach [12], [13] has been adopted. We will now review the mechanism for integrating Copy-Paste checking and design of the similarity checking services.

4. Mechanism for Integrating Copy-Paste Checking

Any activity in a learning scenario in ICARE can be attached to an event that could be tracked by the system, in order to invoke the Copy-Paste manager service. Figure 5 illustrates the processes involved in the integrated copy-paste checking.

The Copy-Paste manager requests a similarity checking service to perform similarity detection on a selected database(s). The results of this check are then passed on to an aggregator service to compile its findings, which is the forwarded to the copy-paste manager. The aggregator is able to compile the results of a group of students and present statistical information as required for the activity. The Copy-Paste manager informs the associated learning object which then provides feedback to the students. It will also be responsible for preparing a report for the instructor and mentor(s).

Figure 5: Design of Integrated Copy-Paste Checking

An experimental copy-paste checking feature is being incorporated into WBT-Master. An instructor is able to define the form of copy-paste checking to be integrated to a selected learning activity. The interface for defining the checking as applied for an information supply application is shown in Figure 6.

The key component that supports information supply is the similarity checking facility. The similarity detection capability has been provided as a web service. We will now explore how this service is further applied to perform Copy-Paste checking.
5. Similarity Detection Service for Copy-Paste Checking

Text similarity is a basic function that determines the degree of similarity between a text to be evaluated and a document(s) is found either on the Web or in an internal student database. This system makes use of an enhanced conceptual text similarity detection approach. As opposed to traditional plagiarism detection tools, which check for copying when students submit their work, our approach is unique in that an instructor can verify student inputs at any point in their learning activity and request a service to compute a score that determines the ability of students in carrying out a task. This system also allows the ability to provide students with an immediate feedback to inform him or her of a rightful means of performing an academic task [7].

This service takes as input a text that needs to be checked for similarity with either an existing document in the internal collection or the Web. An index is first constructed for each public and private collection to be used for document similarity checking. Text in a source document is first broken down into moderately sized segments called fingerprints. These fingerprints are used as search queries to identify a list of suspected similar document from available search services. The size of snippets of text used as fingerprints is twenty words, which is similar to the size of snippets returned by search engines such as Google and Yahoo. Snippet sizes can however be varied for extracting either coarse-grained or fine-grained source text segments as required. The fingerprints are compared for similarity with suspected documents via the matching performed on text snippets. A report is then presented to highlight the matched fingerprints and the corresponding degree of copy.

Document similarity is then computed based on the extent of matching fingerprints in a target document. A primary suspected document list is first constructed based on matching fingerprints. Further similarity checking is then performed on the primary list at the document level taking into consideration the number and order of matching fingerprints within a document.

Traditional plagiarism techniques employ word based or text hash based similarity checking. On the other hand, the proposed approach normalizes the text first to extract root forms of words. We employed synsets in Wordnet [15], as a means to enable a deeper concept level checking resulting in a reduction in term set representation. The process of Copy-Paste checking is illustrated in Figure 7 as a set of Web services.

Text being compared for similarity is processed using a Part of Speech (POS) tagger to determine syntactical form of each word. The word is later normalized into its most common or generic form based on the WordNet synonym dictionary. They are represented as term vectors, created in a common vocabulary space of segments for text being compared. The similarity between word vectors of normalized text is calculated using angular measure (dot product) of vectors between the queried text and a searched result. The normalized form of texts allows checks based on normalized root concepts in addition to the typical word based matching. Figure 8 shows the process of text similarity checking approach as used in ICARE.

The Copy-Paste detection design as shown in Figure 7 is a distributed application that consists of composite Web application services to search both the
internet and shared document sources. Service Oriented Architecture (SOA) protocol is employed to support and connect a set of heterogeneous applications made up of open components. Collaborative units (Web services) work transparently within a single computing environment. Web services are developed as standardized application component units built on standardized, platform-independent XML to enable a customized user application. The asynchronous operation of the Web services are simulated via the use of a search proxy. Figure 9 then shows the design of similarity detection system with web service components for document discovery and similarity analysis.

Figure 9: Document Search and Copy Checking Process Design

6. Conclusion

The current state of e-learning and unchecked student expressions have called for the focusing on learners with just-in-time learning support. This paper described the architecture for ICARE, with details of its realization as a mashup. It has further described the design aspects of the integration of Copy-Paste checking facility into an e-learning system. The Web Services based design has demonstrated means of overcoming the limitations of current e-learning systems, which delay the checking of infringements until students submit their assignments. The modular approach applied in the design of ICARE allows the novel ideas to become adapted to other e-learning systems beyond WBT-Master. The specification of the overall architecture and design also serves as a basis for the design considerations of future e-learning systems.

7. References


