AC 2012-3114: OPEN SOURCE SOFTWARE ENGINEERING THEORY: INTELLIGENT EDUCATIONAL TOOL INCREASES PLACEMENT OF GRADUATES IN STEM-RELATED CAREERS

Dr. Naser El-Bathy P.E., North Carolina A&T State University

Naser El-Bathy is an Assistant Professor of electronics, computer, and information technology at North Carolina A&T State University. He earned his B.S. degree from Wayne State University, Mich., M.S, (computer science, 2006) from Wayne State University, and Ph.D. (information technology, 2010) from Lawrence Technological University. El-Bathy is currently teaching at the North Carolina A&T State University. His interests are in health informatics, bioinformatics, artificial intelligence, intelligent information retrieval, and intelligent web development. El-Bathy may be reached at nielbath@ncat.edu.

Dr. Clay Samuel Gloster Jr., North Carolina A&T State University
Dr. Ghassan M. Azar
Dr. Cameron Seay, North Carolina A&T State Univeristy

Cameron Seay has a B.A., City University of New York, a M.A., State University/Albany, N.Y., and a M.S., M.B.A., and Ph.D., Georgia State University.

Mr. Mohammed K. El-Bathy, Lawrence Technological University

Mohammed El-Bathy joined Lawrence Technological University at 2000. He taught undergraduate and graduate courses including: Computer Science 2, Computer Architecture and Assembly Language, Comparative Programming Languages, Operating Systems, Introduction to Distributed Computing, Advanced Distributed Computing and Distributed Database Systems. Prior to teaching at LTU, El-Bathy was an Adjunct Professor at Wayne State University where he has taught Operating Systems, Comparative Programming Languages and Data Structures, and Algorithms. El-Bathy's area of research is distributed computing and computer networking. He is interested in applying the concepts and techniques of distributed computing, information extraction, intelligent information retrieval, and data warehousing using service-oriented architecture (SOA) to develop intelligent web applications. El-Bathy received a B.Sc in business administration from Cairo University in Egypt and a post-bachelor's degree and M.S. in computer science from Wayne State University. He also has completed all computer science courses and passed the proficiency exam of the Ph.D. program at Wayne State University. In addition to his academic experience, El-Bathy has more than 25 years of experience in industry, working in areas including software development, database design, and computer networking design, implementing projects across many technology platforms, DBMS, network topologies, and programming languages. As a practitioner of information technology, he reached the pinnacle of his career and held a Vice President position of the Department of Information Services for a $1.6 billion financial institution, providing leadership in applying and aligning information technology with business strategies, goals, and needs. While the study of and research in computer science is his prime objective, his interests in literature, music, travel, and nature help him to maintain a sense of perspective in life. He likes to write and has published some articles and is a co-author of a published book. He believes that each of us must give something back to society, so he contributes to local organizations that focus on the environment of his hometown.

Dr. Ibraheem A. Kateeb, North Carolina A&T State University

Ibraheem Kateeb received his B.S. in physics and mathematics from Yarmouk University in Jordan, and M.S.E.E. and Ph.D. from NCA&TSU in North Carolina in electrical and computer engineering. He is a Senior Member of IEEE and Chairman of CNC-IEEE with more than 20 years of experience in academia and industry. He was professor and Department Head of Electronics Engineering at Guilford Technology College. He is currently at NCA&TSU as Assistant Professor of electronics, computer, and information technology. His current research is on electronic components, green energy and power, and control-robotics. He has more than 20 journal, book chapters, and peer-review publications in these areas.

Dr. Rajeev K. Agrawal, North Carolina A&T State University

Rajeev Agrawal is an Assistant Professor at North Carolina A&T State University. His research interests include cloud computing, network security, and content-based image retrieval.

Mr. Aiman Ghassan Baset

Open Source Software Engineering Theory:

Intelligent Educational Tool and Research Methodology

Abstract

The development of World Wide Web (WWW) a little more than a decade ago has caused an information explosion that needs an Intelligent Web (IW) for users to easily control their information and commercial needs. Therefore, engineering schools have offered a variety of IW courses to cultivate hands-on experience and training for industrial systems. In this study, Open Source Software Engineering Theory (OSSET) project course has been designed to help students learn theoretical concepts of IW, practice advanced technical skills, and discover knowledge to solve problem. Undergraduate Science, Technology, Engineering and Mathematics (STEM) students involved in the development of innovative approaches and techniques. They are able to help solve the problems of disease misdiagnoses that medical and healthcare professionals experience. They co-authored and presented numerous research papers introducing the solution in different conferences. This study provides the solution in the form of an Intelligent OSSET using Service-Oriented Architecture (SOA) to decrease disease misdiagnosis in healthcare.

The proposed project course has become a way to establish an “Intelligent Open Source Software Engineering for Healthcare IT" center in our department. Results show that this new course strengthens the capacity and quality of STEM undergraduate degree programs and the number of overall graduate student enrollment. It promotes a vigorous STEM academic environment and increases the number of students entering STEM careers. It expands the breadth of faculty and student involvement in research and development. It enhances and leverages the active engagement of faculty technology transfer and translational research. It improves and develops new relationships between educational institutions and research funding entities to broaden the university's research portfolio and increase funding. The proposed project course is a software engineering research methodology, an educational tool, and a teaching technique is needed in future medical and health IT fields.

Introduction

Last decade, the researchers have designed and developed several intelligent web technologies such as Web Mining (WM) and Web Services (WS). These technologies have become the major courses that provide engineering graduate students with intelligent web skills. Some schools offer these courses as elective courses in undergraduate program. Others recommend it as directed study courses for undergraduate and graduate students. OSSET research project has been evolved as a key course at North Carolina Agricultural And Technological State University, and Lawrence Technological University in the State of Michigan. In the fall of 2010, the course has been offered for the first time at Lawrence Technological University as a directed study course for undergraduate program. This research project prepares students for STEM careers using the criteria of Service-Oriented Architecture (SOA), Artificial Intelligent, Bioinformatics, Intelligent Information Retrieval, Web Middleware, and Server Technologies.

El-Bathy designed the course as a software engineering research methodology, an educational tool, and a teaching technique. As a research methodology, the instructor addresses the conceptual aspects of innovation and discusses the research complications associated with the notion. The instructor also outlines a list of factors said to contribute to innovation within organizations. The course is an educational tool that the instructor uses in teaching an array of technologies. This tool is an extensive workshop in which the students learn these new technologies, implement it, and discover knowledge to solve problems using technical skills they learn. The teaching technique is a structure in which the development of the research project is formed, designed, and managed. This technique enforces the concept of software engineering. It ensures accuracy, efficiency, and high quality during the process of the research project analysis, design, assessment, implementation, test, maintenance and reengineering.

Web Information Retrieval (IR) courses are being offered for both undergraduate and graduate students in many schools such University of Arkansas, University of Texas at Austin, New York University, and Lehigh University. Harding University offers Search Engine Development as an elective undergraduate course for sophomores, juniors, and seniors. The course builds a search engine through a set of bottom-up projects. It also develops projects to modify an existing open source search engine.

Motivation

Researchers have often studied open source software engineering solutions for healthcare information technology including OSCAR, FreeMed, TORCH, and OpenEMR. These solutions have provided high-quality electronic medical records, practice management systems, simpler prescription writing, scheduling, and billing. However, the authors believe that these solutions cannot entirely solve the problems of disease misdiagnosis because of its incapability to check diagnoses with symptoms. Motivated by these problems, the authors propose "Open Source Software Engineering Theory: Intelligent Educational Tool Increases Placement of Graduates in STEM Related Careers". The proposed theory is an automated solution to capture the challenge of disease misdiagnosis while students learn theoretical concepts and technical skills.

The consequences of disease misdiagnosis include unnecessary treatments and testing, long term stay for the patient, high costs and major health risks, useless resources, lateness, and unreliability. The causes for this challenge involve four main factors: absence of open software systems' integrity, inefficient information retrieval processes, poor quality of clustering different diseases' relevant information, and lack of information that analysts require to strategically plan medical and healthcare industries.

Course philosophy

The philosophy of this course project is based on its level. In an undergraduate program, an introduction to intelligent web development course is designed and structured. The course is highly motivated forward looking students in computer science, engineering, education, instructional technology, medical science, and management. After completing this course the student are acquainted with fundamentals of Service Oriented-Architecture (SOA), XML schema, fundamentals of Semantic Web, introduction to Artificial Intelligence, Search Methodologies, Service Orchestrations with Business Process Execution Language (BPEL), Introduction to Web Applications development, and Introduction to IT Research Methodology.

In a graduate level, advanced intelligent web development course is designed and structured. The course is of interest to graduate students in computer science, engineering, education, instructional technology, medical science, and management. Students master new technologies such as: Business Process Execution Language, Java Server Faces (JSF), Web Services, SOAP, WSDL, UDDI, APIs and XML. In this course, we use major platforms for web application and web services development such as Oracle Server Application (OSA) and Java EE Application server, along with IDEs such as JDeveloper. All background material related to HTML, XML, JavaScript, Java SE/EE, and client/server architecture are developed within the course itself from scratch. The course is for students who prefer hands on experience of advanced IT Applications and research methodologies and like the thought of using real tools. It is also for students who want to be graded based on what they can do as well as what they know and the students who are interested in writing, publishing, and presenting papers in scientific conferences and journals.

The Software used in the course includes:

- Design Tool: MS Office Visio Professional
- DBMS: Oracle
- Java: jdk-1_6_0-rc-windows-i586.exe
- Web Server: Oracle Server Application
- IDE: Oracle JDeveloper
- JDBC: classes12.zip

Thus, Open Source Software Engineering Theory (OSSET) project course is an integration of theory and practice approaches. This paper focuses on the discussion of these approaches by providing a technical solution that can help in solving the problems of disease misdiagnosis in healthcare.

The instructor introduced concepts and approaches of technologies, techniques, and software tools that are needed to complete the project. The objective is to get students to be familiar with these concepts to develop the course project. The instructor divided the class into teams. Each team member had a primary task with his/her team and a secondary task with other teams. Each team selected a team leader. The role of team leaders was assigning a task to each team member, clarifying the procedures of each task, solving problems, and providing a weekly progress report to the project manager, the instructor. The tasks are based on Software Development Life-Cycle (SDLC) phases. These phases are planning, implementation, testing, documenting, Deployment, and maintenance. The students trained on each of these phases.

At the same time, the instructor initiated IT Research Methodology that the students followed during the development of the project. The instructor presented research concepts and approaches. These include research purpose and process, research classifications, Institutional Review Board (IRB), scientific research approach, innovation, research process model, research methodology, and research criteria.

The remainder of the paper presents the developments of the students including a new intelligent clustering based extended Genetic Algorithm (ICEGA) using Service-Oriented Architecture, a discussion of research challenges for two main components of ICEGA: data accuracy with Service-Oriented Architecture principles and the prototype that validates the research, preliminary results, and discussion of related work.

Open Source Software Engineering Theory (OSSET) project

Automated clustering of information relies on the ability to programmatically adapt over time to find new methodologies necessary to break data into meaningful clusters. With data constantly changing, it is desired to develop an algorithm capable of clustering in a way that is relevant to the data that is being clustered. In order to tackle this problem, the algorithm must have the ability to try numerous ways of clustering a particular data set.

In an attempt to allow for this capability, the use of an intelligent clustering based extended genetic algorithm has been put in place to provide a way of clustering data that is relevant to the type of data being clustered, with the ability to adapt over time to changes in subjects of topics of desired data. By developing such algorithm, data can evolve into information in a way that produces robust flexibility.

Researchers have often studied general algorithms and technical types of information systems which cannot entirely solve these problems. Therefore, the authors claim that the industries' organizations still face severe obstacles mainly in clustering relevant information that have adapted over time. This claim is derived from the observation of the results of disease misdiagnosis in medical and healthcare industries. Such results include unnecessary treatments and testing, long term stay for the patient, high costs and major health risks, useless resources, lateness, and unreliability. The incidence rate of misdiagnosis is rationally ranges from 1.4% in cancer biopsies to a high 20-40% misdiagnosis rate in emergency or ICU care. Patients' surveys show that diseases misdiagnosis ranges from 8% to 40%. The rate of "failure to diagnose and treat in time", most common reason for a patient safety incident, is 155 per 1,000 hospitalized patients.

Current research has improved data clustering by applying different algorithms to group diseases according to patient's symptoms. However, the authors claim that even if these algorithms can find a solution faster, the quality of data clustering and relevancy between symptom-matching and relevant diseases remain a challenging research problem.

In this paper, the problem of clustering intelligent web search engine using K-means algorithm has been analyzed and the need for a new data clustering algorithm such as Intelligent Clustering Based Extended Genetic Algorithm (ICEGA) is justified to improve the process of disease diagnosis. While K-means is useful and efficient when it comes to clustering data, it lacks the ability to intelligently evolve over time to user browsing patterns and collected data topics. In this paper, the concept of genetic algorithm based clustering has been modified and applied to provide better diseases clustering results in a more efficient manner.

To our knowledge, this work is the first optimal approach for clustering based extended genetic algorithm. ICEGA is a complementary research. It does not disqualify current information retrieval and data clustering research. The goal of ICEGA is to address the applicability of

potential extended genetic algorithm to solve the efficiency and limitation problems in data clustering. To achieve this goal, this course project integrated concepts and approaches of search methodologies, information extraction, intelligent information retrieval, clustering, extended genetic algorithm, and data warehousing. This project is designed and developed in a SOA environment to enable an intelligent architecture.

In this paper, the authors examined a fundamental theory for ICEGA that can establish the groundwork for more future research. This theory is a new attempt to apply SOA principles by providing dynamic services that have concrete meaning on the industries level to improve the capability of the organizations. These services enable Intelligent Information Retrieval Lifecycle Architecture as a requirement to help solve the problems of clustering relevant data with the ability to adapt over time.

A prototype is created and examined in order to validate the concepts. This project involves collaboration with domain scientist and students to evaluate ICEGA on important scientific computing application. Also, the authors collaborate with the Children's Hospital of Philadelphia to increase the number of students and underrepresented cultural minorities in undergraduate research.

Intelligent clustering based extended genetic algorithm

Genetic algorithm is considered to be one of robust and efficient search and optimization technique that was inspired by evolutionary biology and computation research. Traditionally, GA uses fixed-length bit string of natural selection of living organisms for representation.

In our project, we proposed ICEGA mechanism to be an optimal solution for data clustering to improve the efficiency and performance for retrieving a proper information results that satisfy our user’s needs. ICEGA can use several mutation operators simultaneously to produce next generation. This series of random mutation process depend on chromosome best fitness in the population and also rely on high relevancy as well. The mutation operation guarantees the success of genetic algorithms for data clustering since it expands the search. So the highly effective mutation operators the greater effects on the genetic process. Finally, The ICEGA for data clustering gives the user needed documents based on similarity between query matching and relevant document mechanism.

Data Preparation and Clustering

The purpose of our clustering algorithm is to divide set of N documents into K clusters, where the sum of distances D between clusters' documents is the least possible. This means that when clustering algorithm has been completed, the set will be divided into K proper subsets with no documents in more than one such subset of the documents. Each subset has the closest grouping of documents possible with K clusters.

In our clustering algorithm, each document is stored both as a set of weights and a set of words that the weights correspond to. The set of weights is the ratio of each word's occurrences to the sum of all words in the document's occurrences. To simplify some of the computations involved, each document's set of words contains every word that appears in any of the other documents, but with a weight of zero if it does not actually occur within that document. Euclidean distance is

utilized in computing the similarity to quantify the distance between the documents in each cluster. The average of the distances between all documents in each cluster to each other, as if they were points in an n-dimensional space is used as our "quality" for each cluster. In an n-dimensional space, n is the number of words in each document.

The following math is used to find D, the average distance between the documents in the i_{th} cluster of set C of clusters.

$$P = C_i \times C_i \tag{1}$$

$$d(A,B) = \sqrt{(A_1 - B_1)^2 + (A_2 - B_2)^2 + \ldots + (A_{|A|} - B_{|B|})^2} \tag{2}$$

$$D = \left(\frac{1}{|P|}\right) \sum_{i=1}^{|P|} d(P_{i_0}, P_{i_1}) \tag{3}$$

The variable P is used to hold the Cartesian product of the set of documents in the cluster with itself, creating a set of pairs of documents. Each pair in P contains two documents from within the cluster, and to find the average distance between any two documents in the cluster, each pair's distance will need to be computed. The function d is the Euclidean distance between two sets. D, the average distance between the documents in C_i, is calculated by finding the sum of all distances of P's elements and finding the quotient of that and the cardinality of P.

In this paper, the structure of genetic algorithm is extended to hold multiple populations in the population space. The ICEGA is designed using artificial intelligence methodologies, not geometric approaches, to the clustering problem. Our proposed method uses a genetic algorithm to find an ideal clustering solution instead of a more mathematical method such as the K-means algorithm. This key difference allows for more adaptive behavior within our clustering method.

This paper builds a utility-based intelligent agent that implements a faster genetic algorithm with greater efficiency than the original algorithm. The clustering process involves a series of mutations that will evolve over time taking only mutations with a high relevancy, and mutating those further. Figure 1 describes Intelligent clustering based extended Genetic Algorithm (ICEGA).

1. Create initial random population P of N individuals
2. i ← 0
3. If i is equal to the number of desired generations,
   return the best individual of the most recent generatic
4. P_{i+1} ← empty set
5. B ← the most fit individual of the previous generatic
6. Add B to P_{i+1}
7. Insert into P_{i+1} mutate(B)
8. Repeat 7 until P_{i+1} has N individuals
9. Evaluate the fitness for all individuals from P
10. i ← i + 1
11. Goto 3

Figure 1 The Algorithm

Fitness

The fitness of an individual is computed based on the “distances” between the words or other tracked items appearing within a document. The items are compared by their weights, meaning the ratio of their appearances to the total sum of words in the document. These weights are then treated as if they were coordinated for the document's point on an n-dimensional grid, where n is the number of different words appearing within the set of documents being clustered by ICEGA algorithm.

In ICEGA algorithm, an individual with a lower fitness value actually represents a solution of greater quality than one with a greater fitness value. This is because the quality of the clustering solution is the closeness of the items being clustered. Only the most individual fit is passed on to the next generation. The fitness for a chromosome is found through repetition of the math used for finding the similarity of the documents in a cluster. For each chromosome in the generation, the fitness is computed by finding the average of the similarities for each cluster. By using this method, the fitness is also the average distance between any two documents in any one cluster in the solution.

Mutation

Mutation is a way that changes the population to produce the best solution. The ICEGA clustering process involves a series of mutations that will evolve over time taking only the mutations with a high relevancy, and mutating those further. The ICEGA algorithm used one type of mutation. This type is known as a one-point mutation. A single document's position is moved through the chromosome, switching its place in the clusters with another document. Through the repeated use of this type of mutations, the solution can create a generation consisting of a multitude of clustering possibilities.

To further increase the genetic diversity present in each generation of the ICEGA, the algorithm includes a step where a new individual is added to the population. This individual is randomly generated with each generation iterated, to create additional diversity, even without the crossover step's inclusion in the algorithm.

Crossover

The proposed algorithm would build new chromosomes out of sections from two different chromosomes, creating new generations with greater diversity. The lesser number of generations required comes with a cost in the form of a drop in efficiency.

Chromosomes are encoded to represent a genetic algorithm and to be parsed into tree structures. Currently, our genetic algorithm stores each chromosome as a sequence of characters representing the documents. The order of the characters in our chromosomes is of great importance and no repeats are allowed. Using crossovers in the source code of our genetic algorithm negatively affects the efficiency of the algorithm more than it would lower the amount of generations required. The proposed genetic algorithm is simply a way to go through a vast number of possible solutions with greater speed and efficiency than other strategies. With or without crossovers, our genetic algorithm should arrive at the same value.

Research challenge 1 - Data accuracy with SOA

As it is important to manipulate data accurately and efficiently, Service-Oriented Architecture approach has been proposed. Because SOA is a growing successful paradigm, it enables the development of this project as smoothly integrated and reused web services. The benefits of using SOA include reduction of development time and integration costs. Therefore, Service-Oriented Architecture is a central part of the concept that is proposed in this project. It implements dynamic service capabilities with intelligent clustering based extended genetic algorithm to apply reasoning and flexible service workflows.

As the research focuses on the development of intelligent clustering based extended genetic algorithm using service-oriented architecture, it introduces intelligent information retrieval lifecycle architecture with the ability to adapt over time to changes in subjects of topics of desired data. Figure 2 describes the architecture.

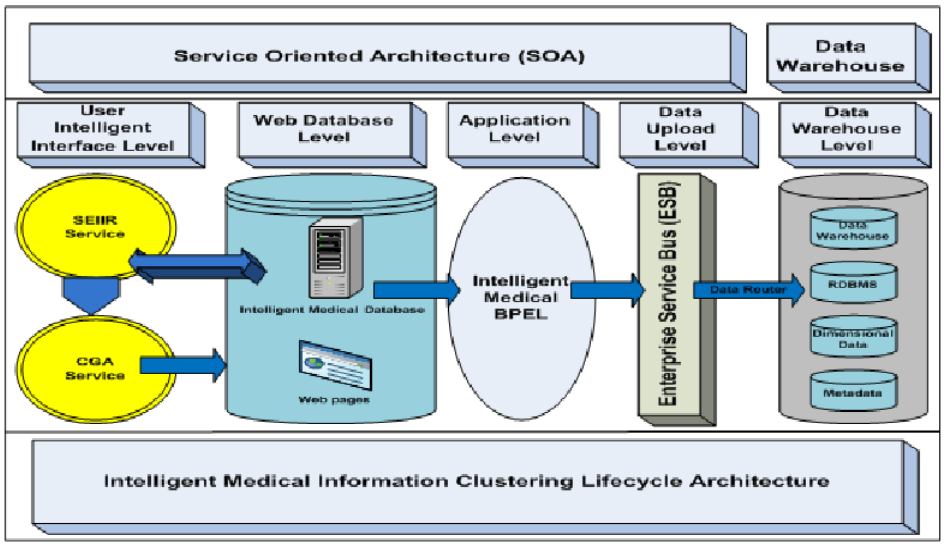

Figure 2 Information Clustering Lifecycle Architecture Based Extended Genetic Algorithm using SOA

One specific research question which arises is: How does the integration of search methodologies, intelligent information retrieval, intelligent clustering, extended genetic algorithm, and intelligent agents using SOA solve the efficiency and limitation problems in data clustering? In the course project, the students deployed SOA middleware as a suite consisting of:

Web service

A web service is a technology that enables programs to communicate through Hypertext Transfer Protocol (HTTP) on the Internet. The students published and consumed two web services to perform operations that are required for developing the project. The services' operations include:

Search, extract, intelligent information retrieval (SEIIR) web service

The first operation is Search Engine (SE) that searches web and local databases for a query string. The second operation is Information Extraction (IE) that extracts text from the source code of web documents. The third operation is Intelligent Information Retrieval (IIR) that

retrieves top ranked documents that are relevant to query strings. This operation involves document/query representation, document ranking, retrieval modeling, and retrieval quality evaluation.

Intelligent Clustering Based Extended Genetic Algorithm (ICEGA) Web Service

This service performs operations that are needed for clustering top ranked documents/diseases. Once ICEGA algorithm is put in place, the desired service item can be requested. Upon this initial request, the first generation of information retrieval is randomly generated, which can lead to a slight decrease of efficiency. What makes up for this initial sacrifice in performance is that as the workflow processes information, the algorithm creates a new generation of logic and the results are assessed based on goodness of fit to results. As new logic workflows are developed, they can be selected and mutated to produce better results. As this process continues, eventually the operation IIR can be provided to matchmaking with user requirements in such a way to enable increased efficiencies over time. Upon delivery of the user request, the generation cycle is terminated.

Business Process Execution Language (BPEL)

The orchestration of web services is supported by Business Process Execution Language (BPEL). In this course project, the students simply designed, deployed, monitored, and administered the process within a framework provided by Oracle BPEL Process Manager. BPEL enables linking SEIIR and ICEGA services as one piece of a process.

Enterprise Service Bus (ESB)

ESB is the services' loosely coupled groundwork utilizing SOA for providing improved business flexibility, reusability, and largely reaction in message-oriented environment applying industry standards. In this research, the students implemented ESB to transform and rout intelligent information from operational database to data warehouse.

Oracle Application Service (OAS)

OAS is standards-based software system server. It enables complete platform integration for executing SEIIR, ICEGA, and Intelligent BPEL process. The students deployed, executed and tested using OAS.

Research challenge 2 - prototype model

The prototype of the research is a simulation of the conceptual solution which can be applied in a real world. The students applied Architected Rapid Application Development (ARAD) prototype model. The prototype intelligent processes are Information Retrieval (IR) and Clustering Extended Genetic Algorithm (CEGA).

Prototype projects

The students developed three types of projects. Projects that provide services. These services are SEIIR (Search, Extraction, and Information Retrieval) and CEGA (Clustering Extended Genetic Algorithm). IIRLABPEL project that defines flow of action in the application. It invokes projects

that provide services. A web front-end application called the IIRLAUserInterface is provided such that the system can be invoked by the users.

The projects are invoked in the following order. When a user enters a query string using the IIRLAUserInterface application, this action invokes the IIRLABPEL project. The IIRLABPEL project defines the main flow of the system. The SEIIR project receives the query string and returns query ids. The CEGA project clusters the documents and writes document's ID, the query's ID, and the cluster name to the database.

Technologies and Techniques

The students integrated SOA Suite technologies such as BPEL to invoke web services in a defined flow sequence. Table 1 lists the technologies and techniques used in the projects. The requirements of the prototype's system are translated into an object data model. The model is transformed into object class databases that store the data. Figure 3 illustrates the model.

Table 1 Technologies and Techniques Used in Each Project

Technologies	Web Services	Tables	Techniques
IMEIRLAUserInterface			Shows how to invoke the ISLABPEL project from the "Search" button.
IMEIRLABPEL	SEIIR ICEGA		Shows how to use BPEL to orchestrate a flow sequence. Invokes the services provided by all the projects
IMEIRLAWS	SEIIR ICEGA	query DocInfoExtraction stopwords ClusteringGA	Shows "bottom-up" implementation of web services: starting with Java classes, you use JDeveloper to generate a WSDL file. Uses JDBC new internal method

In the course project, a real-time data warehouse using SOA is designed. Variable data of different bundled database systems are obtained and captured by a web service.

Figure 3 Object Class Data Model

Prototype Walkthrough

The techniques of walkthrough are approved as experimental assessment approaches to evaluate system application usability. The course project carried out a contextualized usability assessment walkthrough technique that examines the prototype. The walkthrough method evaluates the different phases of the research process. During the system evaluation phase, the examiners evaluated the interfaces that are related to real roles and real users.

The walkthrough examiners of this study are professors, researchers, and SOA engineers in North Carolina A&T State University. They identified different types of problems. These types include design, development, testing, usability, and maintenance problems. They verified that the prototype satisfies the requirements of this research. Also, the prototype is evidence that proves the new concept is valid, the solution is conceptualized, and the findings answered the research question and solved the research problem.

Preliminary results

The ICEGA algorithm is tested on set of sample data. The data is based on 50 generations/iterations of the ICEGA or K-means respectively, using the same random sample set of 15 documents with 600 words each. Figure 4 serves as decent evidence that the solutions from our Intelligent Clustering Based Extended Genetic Algorithm are generally closer clustered than those generated by K-means, even if K-means can find a solution faster. Figure 4 defines GA 1 and GA 2 as the two graphed trials of the genetic algorithm.

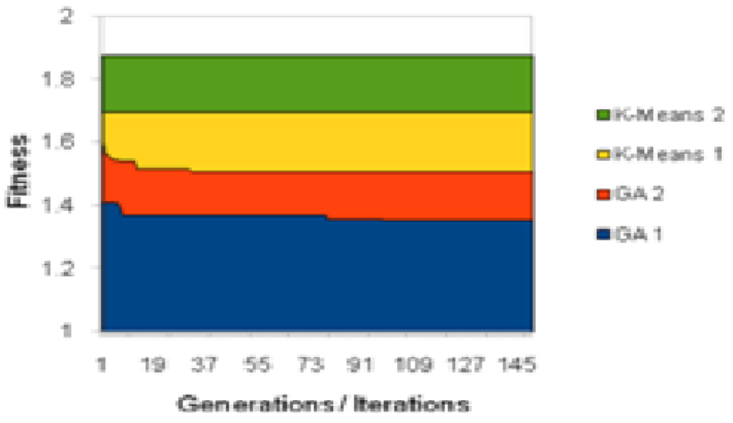

Figure 4 ICEGA and K-Means Comparison

Figure 5 presents sample set of 15 documents as a demonstration of clustering. The document set has been simplified to only have 2 different words in each document. The values on the X and Y axes are the word weights of those two words in the documents. Figure 5a shows the documents arranged on 2-dimensional grid without any clustering information applied. Figure 5b and 5c differ in that the documents have been colored and circled to designate the different clusters within the set of documents. Figure 5b has been clustered using the K-means algorithm, while with Figure 5c our genetic algorithm is used to find a clustering solution.

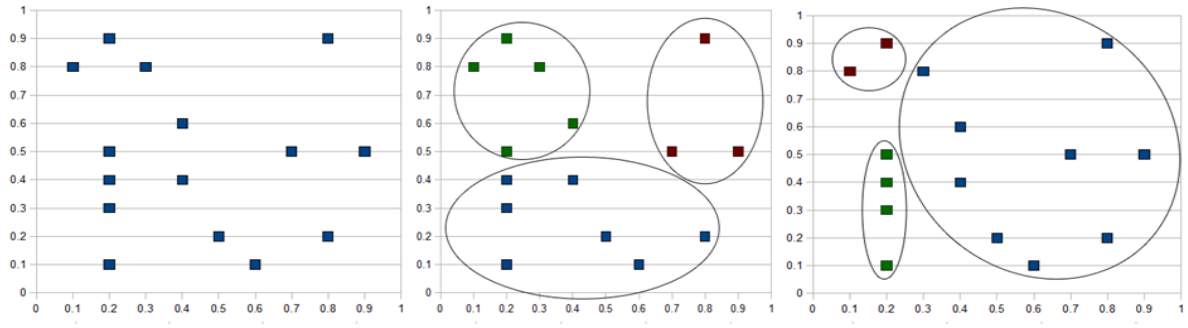

Figure 5 a) Documents without clustering (left), b) K-means Clustering Results (middle), c) ICEGA Results (right)

The results are listed in Table 2 were collected over 15 test runs of both clustering methods on the same data set. The table shows the statistics collected from ICEGA and K-means algorithms to demonstrate their relative performance capabilities. The values given are the fitness of the final clustering solution generated by each run, which means that the lower fitness are from better solutions, while higher fitness values are worse solutions. As each method uses a random starting point, there is room for variation in solutions.

From this data, we can observe that on average, our ICEGA algorithm excels K-means clustering algorithm. The test runs did not find as good a solution with K-means as the best solution from the ICEGA algorithm, and even the worst solution from the ICEGA algorithm is of better fitness than the average solution from K-means.

While the data collected does not represent all possible input cases, and cannot claim to represent all of them, it shows a trend of the ICEGA algorithm exceeding the performance shown the clustering process we had used previously.

Table 2 ICEGA and K-means Performance

	ICEGA	***K-means***
Maximum	1.66384	1.86476
Average	1.56938	1.67881
Minimum	1.35574	1.40269

The preliminary results show that the proposed algorithm outperforms K-means algorithm. The proposed concept ensures high level of accuracy and efficiency due to removal of irrelevant information. The Clustering Intelligent Extended Genetic Algorithm (ICEGA) enhances an organization's ability to collect information faster at lower cost and to make accurate decisions. The orchestrations of clustering extended genetic algorithm by applying SOA principles and concepts allow flexible service workflows to be immediately adjusted to modifications and make systems smarter. Preliminary results also show that ICEGA can discover related diseases to doctors' original diagnosis and automatically reassesses the situation if their diagnosis is incorrect. The proposed algorithm solution markedly increase the success of disease clustering and relevancy between patient's symptoms and diseases.

In addition, the instructor asked the students to complete a job survey and return it once they obtain a job in any of the areas that they worked on during the course the project. Figure 6 shows

the number of jobs that offered to students in each of the skills learned in course project in the past two years.

In web service technology, 160 students received job offer. In SQL and XML, 150 students received job offer. In SOA and BPEL, 140 students received job offer. In Java, 120 students received job offer.

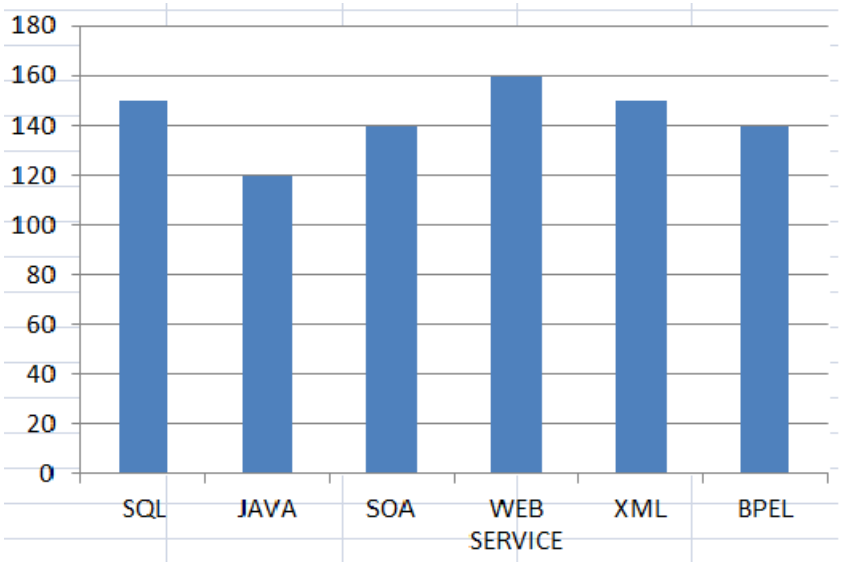

Figure 6 Number of jobs offered to the student

Related work

Previous work in data clustering has focused on concepts similar to Intelligent Clustering Based Extended Genetic Algorithm. K-means is most successfully used on data sets because of its simplicity and its linear time complexity. However, it is not feasible to be used on large data sets. Hierarchal clustering algorithm creates a structure that reflects the order of divided groups. It gives better results than K-means if it uses random data set. A GA-based unsupervised clustering technique selects cluster centers directly from the data set and allows acceleration of the fitness evaluation via a look-up table. A limitation of existing techniques is the inability to adapt over time to changes in data. Such techniques do not provide a general architecture that enables any operation to be automatically optimized for any system.

Conclusion

Open Source Software Engineering Theory (OSSET) project course is a software engineering research methodology, an educational tool, and a teaching technique. It also helps students learn theoretical concepts, practice advanced technical skills, and discover knowledge to solve problem. The course satisfies the needs of undergraduate and graduate students in computer science, engineering, education, instructional technology, medical science, and management. This new course strengthens the capacity and quality of STEM undergraduate degree programs and the number of overall graduate student enrollment. It promotes a dynamic STEM academic environment and increases the number of students entering STEM careers.

Acknowledgements

The primary author of this paper, Dr. Naser El-Bathy, gratefully acknowledges the students who enrolled in this course project for their significant contributions to achieve the goal, objectives, and activities of this research.

Bibliography

[1] R. H. Abrahiem. A New generation of middleware solution for a near-real-time data warehousing architecture. *Electro/Information Technology IEEE International Conference*, pp. 192–197, May 2007.

[2] F. Gomez, B. Chandrasekaran. Knowledge organization and distribution for medical diagnosis. *IEEE Transactions on Systems, Man, and Cybernetics*, Vol. SMC-I 1, No. 1, Jan. 1981.

[3] R. T. Watson. Data Management: Databases and Organizatins. *Wiley*, 2006.

[4] Q. Gu, P. Lago. A stakeholder-driven service life cycle model for SOA. *ACM, New York*, pp. 1-7, 2007.

[5] P. Kudov´a. Clustering genetic algorithm. *IEEE, DOI. 10.1109/DEXA, 65*, 2007.

[6] B. Coppin. Artificial intelligence illuminated. *Sudbury, Massachusetts: John and Bartlett Publishers*, 2004.

[7] C. Perks, T. Beveridge. Guide to Enterprise IT Architecture. *New York: Springer-Verlag*, 2003.

[8] J. Cresswell. Research Design: Qualitative, Quantitative, and Mixed Methods Approaches. *Thousand Oaks, London, New Delhi: SAGE Publications International Educational and Professional Publisher*, 2003.

[9] D. Remenyi, B. Williams, A. Money, E. Swartz. Doing Research in Business and Management – An Introduction to Process and Method. London. *Thousand Oaks, New Delhi: SAGE Publications*, 2005.

[10] D. T. Sanders, J.A. Hamilton, R. A. MacDonald. Supporting a service-oriented architecture. *Society for Computer Simulation International, San Diego, CA*, pp. 325 – 334, 2008.

[11] D. Krafzig, K. Banke, D. Slame. Enterprise SOA Service-Oriented Architecture Best Practices. *NJ: Prentice Hall*, 2005.

[12] T. Erl, Service-Oriented Architecture Afield Guide to Integrating XML and Web Services. *NJ: Prentice Hall*, 2004.

[13] M. P. Papazoglou, W. Heuvel. Service oriented architectures: approaches, technologies and research issues. *The VLDB Journal, Springer Berlin / Heidelberg*, vol. 16, Number 3, pp. 389–415, 2007.

[14] D. Vladimir, "Development of applications with service-oriented architecture for grid," ACM New York, Vol. 374, 2008.

[15] D. Steiner, "Oracle SOA Suite Quick Start Guide 10g (10.1.3.1.0)," Oracle, pp. 7 – 11,2006.

[16] M. Doernhoefer, "Surfing the net for software engineering notes," ACM SIGSOFT Software Engineering Notes, Volume 30 Number 6, Nov. 2005.

[17] A. Dan, R. Johnson, and A. Arsanjani, "Information as a service: modeling and realization," International Workshop on Systems Development in SOA Environments, IEEE, 2007.

[18] R. Akerkar, P. Lingras. Building an Intelligent Web – Theory and Practice. *Sudbury, Massachusetts: Jones and Bartlett Publishers*, 2008.

[19] N. Chaiyarataiia, A. M. S . Zalzala. Recent developments in evolutionary and genetic algorithms: theory and applications. Genetic Algorithms in Engineering Systems: Innovations and Applications. *GALESIA 97. Second International Conference On (Conf. Publ. No. 446)*, pp. 270 – 277, 1997.

[20] D. E. Rowley, D. G. Rhoades. The cognitive jogthrough: a fast-paced user interface evaluation procedure. *CHI '92 Proceedings of the SIGCHI conference on Human factors in computing systems, ACM*, New York, 1992.

[21] D. Pinelle, C. Gutwin. Groupware walkthrough: Adding context to groupware usability evaluation. *CHI '02 Proceedings of the SIGCHI conference on Human factors in computing systems: Changing our world, changing ourselves, ACM*, New York, 2002.

[22] N. El-Bathy, P. Chang, G. Azar, R. Abrahiem. An Intelligent Search of Lifecycle Architecture for Modern Publishing and Newspaper Industries Using SOA. *IEEE, Electro/Information Technology, Normal*, pp. 1-7, 2010.

[23] N. El-Bathy, M. El-Bathy. Intelligent Lifecycle Architecture of Disease Diagnosis Based Adapted CGA using SOA. *KG. Saarbrücken, Germany: LAP LAMBERT Academic Publishing AG & Co*. 2011.

[24] N. El-Bathy, G. Azar. Intelligent Information Retrieval and Web Mining Architecture Applying Service-Oriented Architecture. *KG. Saarbrücken, Germany: LAP LAMBERT Academic Publishing AG & Co*. 2010.

[25] N. El-Bathy, G. Azar, M. El-Bathy, G. Stein. Intelligent Extended Clustering Genetic Algorithm. *Proceeding of IEEE, Electro/Information Technology Conference, Mankato*, MN, pp. 1 – 5, 2011.

[26] N. El-Bathy, G. Azar, M. El-Bathy, G. Stein. Intelligent Information Retrieval Lifecycle Architecture Based Clustering Genetic Algorithm using SOA for Modern Medical Industries. *Proceeding of IEEE, Electro/Information Technology Conference*, Mankato, MN, pp. 1 – 7, 2011.

[27] N. El-Bathy, C. Gloster, I. Kateeb, G. Stein. Intelligent Extended Clustering Genetic Algorithm for Information Retrieval Using BPEL. *American Journal of Intelligent Systems*, pp. 10 – 14, 2011.

[28] N. El-Bathy, C. Gloster, I. Kateeb, G. Azar. Intelligent Search Lifecycle Architecture for Mass Media Using SOA. *Architecture Research Journal*, 2011.

[29] N. El-Bathy, G. Azar, L. Karadsheh, E. Mansour, S. Alhawari. A Theoretical Framework for Knowledge Management Process: Towards Improving Knowledge Performance. *11th IBIMA Conference on Innovation and Knowledge Management in Twin Track Economies*, 2009.

[30] N. El-Bathy, L. Karadsheh, W. Musa, S. Alhawari. Incorporating Knowledge Management and Risk Management as a single Process. *International Conference of the Global Business Development Institute (GBDI).* October 2008.

[31] RightDiagnosis.com. *http://www.rightdiagnosis.com/intro/common.htm*. Accessed November 2011.