Towards Automated Education Demand-Offer Information Monitoring: the Information Extraction

Peteris Rudzajs
Department of Systems Theory and Design
Riga Technical University
Riga, LATVIA
Peteris.Rudzajs@rtu.lv

***Abstract*— Dynamically changing work environment in knowledge economy causes the changes in knowledge requirements for labor. Therefore it becomes more and more important to be constantly aware of what education is currently demanded and what education is currently offered. The IT solution is vital to process various information sources, extract education information, and provide analysis mechanisms in automated manner. The education information extraction is detailed in this paper in the context of Education demand and offer information monitoring system by providing the workflow for semi-automatic skills extraction from the university course descriptions using developed term suggestion method.**

Keywords- education information; monitoring system; information extraction

I. INTRODUCTION

Rapid economic changes in the knowledge requirements for labor cause a necessity to monitor education demand and offer. Education demand and offer (d/o) can be described in terms of knowledge, skills, or competences required in work environment or obtained in university. For simplicity in this paper the terms "education information" and "skills" are used interchangeably to denote knowledge, skills, or competences. Education d/o monitoring both for university and industry can provide an insight in knowledge, skills, or competences currently demanded/offered in educational and industrial environment. To facilitate the monitoring of education d/o, university and industry should be provided with IT solutions that reduce the large amount of manual work currently necessary to overview, extract, and analyze the information from various education information sources (study and certification course descriptions, job advertisements etc.). In this case, the monitoring system design principles are applicable, as basic functionality of the monitoring systems includes gathering of source information, information processing and analysis to provide decision support information to users of the system.

In the previous study [1] the architecture of the Education d/o information monitoring (EduMON) system had been proposed. The architecture (see Fig. 1) consists of various classes of services and supports the process from retrieving and extracting information from relevant sources to finding correspondence between education information in various sources. Education information when reflected in the information sources is represented using domain terminology. Thus the terms of the information source document related to domain terminology should be identified in the process of the information extraction. In the current study an example of semi-automated information extraction method is presented (method contributes to the development of Information extraction services of EduMON system). Solution is illustrated by terminology extraction from the university course descriptions based on the n-grams of text document, Dice's similarity metric and available Free On-Line Dictionary of Computing (FOLDOC) [2] consisting of more than 14 600 IT-related terms. Method is also applicable to extract terms from job advertisements and other relevant sources reflecting the education information in the area of Information and Communication Technology.

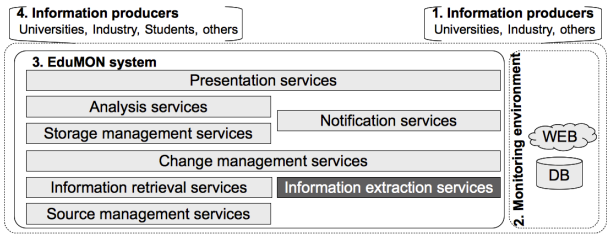

Figure 1. EduMON system's overview

II. EDUCATION INFORMATION EXTRACTION

The extraction of terms related to domain terminology is necessary to provide the standardized reference of terms represented in text. Relevant terms extracted from various information sources can provide basis for applying the analysis mechanisms provided by the EduMON system's services of Analysis service class. Examples of analysis services can include the analysis of skills (represented as terms in information sources) demanded in job advertisements and offered in University courses.

The process of term extraction can be either manual or semi-automatic. In manual extraction, the user manually annotates the document (e.g., University course or job advertisement) by linking the terms presented in document (further in the text - source terms) to the terms available in reference dictionary (further in the text - target terms). As the manual extraction of skills is time-consuming, the solution for semi-automatic extraction is proposed. Basic workflow for the

semi-automatic extraction of skills includes the following processes: 1) the provision of the text section (e.g. a set of paragraphs) representing the skills; 2) the automatic generation of suggestions for the reference of source terms to target terms; 3) the reviewing of suggestions by the user (suggestions for linkage of source terms to target terms can be accepted or rejected; source terms without suggestion can be added as new terms in the reference dictionary); and 4) the storage of relevant pairs of source and target terms.

To automatically provide the suggestions for the reference of source terms to target terms, the education information extraction method TermSuggest is developed. It combines several processes:

1) Preprocessing: Include the conversion of the document text to lowercase and the stopword removal from the text (e.g., removal of such words as "a", "is", "and", "various" etc.).

2) N-gram generation: N-grams, such as 1-gram, 2-gram, 3-gram, 4-gram are generated for the document text.

3) The matching of source terms of the n-gram to target terms based on the selected string similarity metric: The string similarity is calculated between the each source term of the n-gram and the target terms. Any string similarity metric or the combination of multiple metrics can be selected. In the below presented example the Dice's similarity was selected as a string similarity metric. For calculation of similarity metric, the open source Java library SimMetrics [3] was used. The *FOLDOC* [2] was selected as dictionary for target terms. For each source term of the n-gram the best matching target terms are returned as suggestions in the format <*source term, target term*>. As best matches are considered those pairs <*source term, target term*> with the similarity score above the defined thershold.

4) Removal of the irrelevant suggestions: Before the suggestions are provided to the user for the linkage of source terms to target terms, irrelevant suggestions should be removed. As irrelevant suggestions are considered those pairs <*source term, target term*> of the n-gram where the similiraty score is less or equal than maximum similarity of suggestions of the (n+1)-gram where the source term of the n-gram is included in source term of (n+1)-gram.

The TermSuggest method returns the set of suggestions for each n-gram (1-gram, 2-gram, etc.) in the form <*source term, target term, similarity score*>.

To evaluate the results of suggestion, the TermSuggest method had been applied to the example case. Following the workflow of semi-automatic extraction of skills, the document of interest should be selected first, e.g., description of course "Software engineering" is chosen and the section of text to be analyzed is provided. The word count of the provided text section of the document initially is 197 words. Next, the TermSuggest method is applied. The counts of suggestions for each n-gram without and with the removal of irrelevant suggestions for given thresholds 0.65 and 0.80 had been calculated. It had been found that the suggestions with threshold 0.80 are providing more appropriate suggestions for

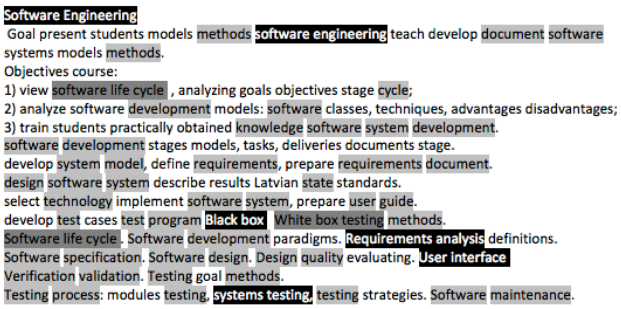
Software Engineering
Goal present students models methods software engineering teach develop document software systems models methods.
Objectives course:
1) view software life cycle , analyzing goals objectives stage cycle;
2) analyze software development models: software classes, techniques, advantages disadvantages;
3) train students practically obtained knowledge software system development.
software development stages models, tasks, deliveries documents stage.
develop system model, define requirements, prepare requirements document.
design software system describe results Latvian state standards.
select technology implement software system, prepare user guide.
develop test cases test program Black box White box testing methods.
Software life cycle. Software development paradigms. Requirements analysis definitions.
Software specification. Software design. Design quality evaluating. User interface
Verification validation. Testing goal methods.
Testing process: modules testing. Systems testing, testing strategies. Software maintenance.

Figure 2. Document text (without stopwords) with source terms matched with target terms with thershold 0.80. Filled text spans represent source terms of the 1-gram (light grey), 2-gram (black), and 3-gram (dark grey) that appear in suggestions.

terms than suggestions with threshold 0.65. It is worth mentioning that for the lower threshold significant number of suggestions was removed thus giving a possibility for user to review more relevant source terms and linked target terms than in the case without applying the removal procedure. Resulting source terms with linked target terms are presented in Fig. 2. Total number of unique target terms is 35.

Following the workflow of semi-automatic extraction of skills, next process is the reviewing of suggestions. The expert evaluated the suggestions and accepted 14 of 28 suggestions for 1-gram, all for 2-gram and 3-gram suggestions. The expert proposed 3 new source terms for addition to reference dictionary, namely, "software system", "software development", and "software design".

III. Conclusions and Future Work

Relevant terms extracted from various information sources can provide basis for applying the analysis mechanisms provided by the EduMON system's services of Analysis service class thus facilitating the monitoring of education d/o. In further work it is necessary to consider the term extraction by taking into account the term dependencies to other terms and the combination of multiple string similarity metrics. The evaluation of proposed method should be conducted on larger document corpus to see its overall performance.

Acknowledgment

This work has been supported by the European Social Fund within the project "Support for the implementation of doctoral studies at Riga Technical University".

References

[1] P. Rudzajs, "Towards Automated Education Demand-Offer Information Monitoring: the System's Architecture," in Selected Papers from Workshops and Doctoral Consortium of the International Conference on Perspectives in Business Informatics Research, BIR2011, 2011, pp. 252-265.

[2] D. Howe, "Free On-Line Dictionary of Computing," 2010. [Online]. Available: http://foldoc.org/.

[3] The University of Sheffield, "SimMetrics," 2011. [Online]. Available: http://sourceforge.net/projects/simmetrics/.