Optimal Automated Method for Collaborative Development of University Curricula

M. Charnine[1] **and V. Protasov**[2]
[1]Institute for Informatics Problems, Russian Academy of Sciences, Moscow, Russia
[2]The Russian Center of Computing for Physics and Technology, Moscow, Russia

Abstract – *Aspects of curriculum design automation using semantic analysis of Web texts to develop candidate keywords and voting alternatives, as well as methods of optimal aggregation of individual expert decisions are presented. It is shown how semantics analysis of online texts can help with: generating of initial version of the curriculum; generating of candidate keywords and alternative options for improving the curriculum; building hierarchy of categories of the subject area or domain; detecting the change of domain knowledge state and defining structural knowledge dependencies; monitoring terminology and knowledge base of the subject area or domain. It is shown how the techniques of optimal integration of individual expert decisions in selecting of automatically generated alternatives helps to create the highest quality curriculum.*

Keywords: curriculum design; semantic analysis of Internet texts, competency assessment, optimal aggregation rules, voting records

1 Introduction

At present, many educational institutions participate in intensive methodological work on development of educational standards, appropriate training programs and curriculums. The high education curriculum must meet the requirements of quality education and follow the speed of technological and scientific revolution boosted by progress in information and communication technologies. At the same time, the study of the semantics of terabytes of Internet texts can help to upgrade the curriculum, making possible fully automated monitoring of the knowledge base.

Next generation of curriculum development systems requires:

- revision of the curriculum development process, including information processing scheme and the use of semantic web searches based on keywords;

- the use of semantic analysis of large volumes of online texts to determine the key terms/keywords of the domain and their semantic relations;

- the use of automated tools to maintain relevance of term/keyword to the specific subject area of curriculum;

- participation of a large number of experts in curriculum development;

- the use of automated tools for collective decision support.

2 Automated tools for curriculum design

The university curriculum is a document which summarizes the structured content, knowledge, skills and abilities for learning by students. The material is usually divided into sections and themes, contains a hierarchy of sections and the sequence of their study.

In many cases, curriculum can be formally presented as a list of keywords/terms of the subject area and a set of hierarchical relationships between them that form a tree-like structure. Such hierarchical structure of the curriculum can be created by a group of experts that have to:

- select keywords/terms for inclusion in the curriculum from a list of automatically generated candidate terms;

- build a hierarchy of selected terms by specifying hierarchical (parent-child) relationships between them.

Experts can use so-called associative portrait of subject area that can be generated automatically to reduce the amount of work on keyword/term selection.

Associative Portrait of Subject Area (APSA) – is a set of associative relationships between keywords/terms that are relevant to the subject area. Automatic methods of finding associative relationships between terms are based on the calculation of their semantic similarity. If semantic similarity between two terms is high then we consider that these terms have an associative relationship. Semantic similarity is also called as semantic distance or connectivity. The method of calculation of semantic similarity is presented below.

2.1 Automatic method of analysis and calculation of semantic similarity

Most automated tools and technologies for curriculum design are based on the method for calculation of semantic connectivity and distance usually called "semantic similarity of terms".

Method of calculation of semantic similarity of terms is based on statistical analysis of large bodies of text from the Internet. In our work the method has been successfully used for keyword research, identification of related and core keywords, calculation of keyword/term clusters, finding categories of clusters and their hierarchical relationships.

The proposed model presents keywords/terms of the curriculum in a form of several clusters of terms that dynamically change over time in quantity, structure and even meaning.

To define, analyze and “calculate” clusters there is a need to understand a measure of semantic similarity between terms also called as semantic distance or connectivity.

The most efficient approach to calculate semantic similarity of terms is based on the use of distributional semantic models [3].

Distributional semantic models were successfully used for the following tasks: word similarity, word clustering, automatic thesaurus generation, word sense disambiguation, query expansion; enabling their application for automated curriculum design.

Distributional semantic models, variously known as vector spaces, semantic spaces, word spaces, corpus-based semantic models, distributional memory, all rely on some version of the distributional hypothesis, stating that semantic similarity between two words/terms can be modeled as a function of the degree of overlap among their linguistic contexts.

In other words distributional hypothesis [3] is claiming that semantically related lexemes/terms have similar context and, conversely, in a similar context, the lexemes are semantically close. The proposed model uses advanced hypothesis, suggesting that context similarities exist not only for semantically similar individual lexemes, but also for semantically similar phrases containing multiple lexemes and terms.

Web contains astronomical and constantly growing amount of textual documents and linguistic contexts, enabling calculation of semantic similarity of terms and curriculum keywords migration with growing precision over time.

The most modern versions of Distributional semantic models are Word Space Models and Distributional Memory Models. Most of the criticism of these models stems from the fact that term-document and word-context matrices typically ignore word order.

In contrast with Word Space Model and Distributional Memory our approach uses more precise method of analysis and calculation of semantic distance and connectivity accounting “fine-tuning” of semantic relations and presence of word combinations and as well as keywords in a context.

Advanced method applied in our research is presented below.

2.2 Advanced method for calculation of semantic similarity

Advanced method for calculation of semantic similarity/distance uses statistical methods for selection of important terms from the body of texts.

These selected terms can be words, or word combination, or named entities. The word order within selected terms is very important and is not ignored.

Let n is the number of selected terms.

Then we create symmetric term–context matrix (n-by-n), which elements are calculated based on co-occurrence of selected terms in a corpus of texts.

When we calculate the matrix elements, we can use special lexico-syntactic patterns (e.g., x "is a" y | y "including" x | x "such as" y) to extract specific types of semantic relationships between terms (hierarchical, genus-species, the part-whole).

Each row of the matrix can be interpreted as a vector in n-dimensional coordinate space corresponding to certain selected term. We call this vector a Context Vector corresponding to certain term.

The set of such context vectors together forms a high dimensional Semantic Vector Space.

Semantic similarity between the terms x and y is calculated based on the cosine measure between the corresponding vectors according to the following formula:

$$\frac{x \cdot y}{|x||y|} = \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x_i^2}\sqrt{\sum_{i=1}^{n} y_i^2}}$$

The terms x and y are considered to be similar/connected/associated if the cosine measure of their vectors has a large value.

2.3 Formal definition of APSA

Formally, the APSA is defined as a graph G = (V, E) with nodes v of V, representing significant keywords/terms and links of the graph (v-i, v-j, Link, w-ij) of E, describe the relationship between terms. Here w-ij is the weight that expresses the measure of semantic similarity between terms. Link represents the type of the context of terms v-i and v-j. This context type is related to type of relationship between terms v-i and v-j. The context type is determined by the parameters of the algorithm for calculation context vectors, such as the size of the context window, or type of syntactic template binding the terms.

2.4 Automatic methods of APSA creation

The process of APSA creation implied a set of methods, including:

-identification of Internet texts related to the specified subject area or domain;

-identification of relevant keywords/terms and their ranking;

-identification and ranking of associative relationships between keywords/terms.

These methods are based on the preliminary studies on texts, including those taken from various Internet resources.

Processing large volumes of texts, constantly updated in the Internet, allows us to collect the necessary statistical data to generate a fairly complete picture of the subject area, presented in the form of APSA. The ability to perform machine learning on a large number of examples gives the system flexibility and improves the results.

Thus automatic nature of APSA creation helps to monitor terminology and knowledge base of the subject area or domain and helps to detect the change of domain knowledge state and define structural knowledge dependencies.

2.5 Method for curriculum development using APSA

APSA technology helps to identify the list of relevant keywords/terms that can be considered as candidate keywords for inclusion in the curriculum. The list of keyword candidates can be also considered as a list of voting alternatives. Experts can vote for every keyword candidate and make collective decision: include it or not in the curriculum.

APSA technology also helps to identify the list of associative relationships between keywords/terms in curriculum. These associative relationships can be considered as candidates for hierarchical relationships between terms. Experts can vote for every candidate relationship and make collective decision: include it or not in the curriculum as hierarchical relationship.

This way experts can create a hierarchy of categories/sections included in the curriculum.

2.6 Algoritm for curriculum development

Formally, the actions of experts to build the initial version of the curriculum and to identify possible changes of its keywords/terms can be described by the following algorithm.

1. Set of initial terms/keywords is given defined manually or taken from the old curriculum

2. Collect text from Internet with predefined keywords (formation of body of texts for various time slots)

3. Design context vectors that determine meaningful closeness and dependency between keywords at various moments of time and time slots (point and evolution dependency)

4. Searching clusters of closely related keywords

5. Searching of the hierarchical structure of cluster dependency (hierarchical clustering).

6. Calculate the center of the cluster of keywords/terms related to the domain or subject area.

7. Those new terms that are closer to the cluster center are candidate terms for inclusion in the curriculum.

8. Those keywords that are more distant from the cluster center are candidate terms for exclusion from

the curriculum. After deletion these candidate terms the cluster center changes its position and thus motion of the curriculum is defined.

3 Automated tools for collaborative decision making

The speed of scientific and technological revolution and the amount of knowledge in every domain is constantly growing. As a result we have a growing number of specific terms in the domain, as well as increasing rate of changes in curriculum terminology.

All this leads to the fact that we need more and more experts working together to maintain up-to-date curriculum.

In a large expert team the diversity of expert opinions and their competences can vary widely and it makes more difficult for them the quick creation of collective decisions. In other words, the larger expert team, the harder natural process of creating timely collective decisions without using special automated tools.

That's why there is a need for automated tool combining individual decisions of experts. This collaborative decision-making (CDM) tool coordinates the functions and features required to arrive at timely collective decisions, enabling all relevant experts to participate in the process. The core output of CDM tool is making better decisions.

This CDM tool can help to make better the following two types of collective decisions:

- whether or not to include a particular term in the curriculum, as well as

- which of the APSA relationships between terms can be considered as hierarchical relationships.

The CDM tool can use Internet for communication purposes and a special optimal algorithm described below for collective decisions support.

4 Algorithm for optimal aggregation of expert's individual decisions

E. Baharad, J. Goldberger, M. Koppel and S. Nitzan in the article "Beyond Condorcet: Optimal aggregation rules using voting records" describe the following optimal algorithm for combining expert decisions.

1) Select some initial weights (shares) for decisions of each expert.

2) For each expert, using the story of his decisions, calculate the probability of convergence of his individual decisions with the weighted collective decision.

3) Calculate the new weighs for each expert based on his probability calculated in step 2.

4) Repeat steps 1-3 until the new weights converge with those calculated in the previous iteration.

Baharad, Goldberger, Koppel and Nitzan indicated that the above algorithm for combining expert decisions works better than any other judgment aggregating rule; in particular, it is better than collective decision making based on a simple majority rule.

The above algorithm uses linear combination of expert decisions. The main result in Nitzan and Paroush (1982), as well as Shapley and Grofman (1984) was the claim that if the probability of a correct decision of each expert is known, then a linear aggregation rule of their decisions is optimal, and the maximum probability of correct collective decision is reached when the weights (shares) of experts are calculated by the formula Wi = log (Pi/1-Pi), where Pi is the probability of a correct decision of the expert with the number i.

Thus, the algorithm described above optimally integrates individual expert decisions into optimal collective decision.

This algorithm provides the highest quality curriculum that can be created by given group of experts with different skills.

5 Conclusions

In this paper we describe the algorithm for the optimal integration of individual expert decisions in selecting automatically generated alternatives, which were taken from the associative portrait of subject area (APSA) that was automatically calculated. The above algorithm provides the highest quality curriculum for a given group of experts with different skills.

6 Acknowledgements

This work was supported by the Russian Foundation for Basic Research, grant #13-07-00958 "Development of the theory and experimental research of a new information technology of self-managed crowdsourcing" and grant #13-07-00272 "The methods for automatic creation of associative portraits of subject domains on the basis of big natural language texts for knowledge extraction systems".

7 References

[1] E.Baharad, J.Goldberger, M.Koppel и S.Nitzan, “Beyond Condorcet: Optimal Aggregation Rules Using Voting Records”, CESifo München, 2011.

[2] A. Lenci, “Distributional semantics in linguistic and cognitive research”, Rivista di Linguistica, 1, 2008, pp.1-30.

[3] M.Baroni, A.Lenci, “Distributional Memory: A General Framework for Corpus-Based Semantics”, Computational Linguistics. V.36, Issue 4, 2010, pp. 673-721.

[4] Peter Turney, “A uniform approach to analogies, synonyms, antonyms and associations”, Proceedings of COLING, Manchester, 2008, pp. 905–912.

[5] M.Charnine, I.P.Kuznetsov, E.B.Kozerenko, “Semantic Navigator for Internet Search”, Proceeding of International Conference on Machine Learning, 27-30, 2005 Las Vegas, USA, CSREA Press, pp 60-65, 2005.

[6] Michael Charnine, Vladimir Charnine. Keywen Category Structure.// Wordclay, USA, 2008, pp.1-60.

[7] Michael Charnine, “Keywen Automated Writing Tools”, Booktango, USA, 2013, ISBN 978-1-46892-205-9.

[8] Nitzan, S., and J. Paroush. 1985. “Collective Decision Making: An Economic Outlook”, Cambridge University Press, Cambridge, England.

[9] Nitzan, S., and J. Paroush. 1982. “Optimal Decision Rules in Uncertain Dichotomous Choice Situations.” International Economic Review 23(2): 289-97.

[10] Shapley, L. and B. Grofman. 1984. "Optimizing Group Judgmental Accuracy in the Presence of Interdependencies." Public Choice 43: 329-343.

[11] R. Diamond, “Designing and improving courses and curricula in higher education: A systematic approach: Jossey-Bass,” SF,1989.

[12] P. Ramsden, “Learning to teach in higher education,” London: Routledge, 1992, DOI: 10.4324/9780203413937.

[13] S. Toohey, “Designing courses for higher education,” Buckingham, UK: The Society for Research into HE & Open University Press, 1999.

[14] M. Robbins, I. Schagaev, J. Yip, “Effective teaching in theoretical IT modules,” 33rd Annual Frontiers in Education Conference," (FIE'03), 2003, http://doi.ieeecomputersociety.org/10.1109/FIE.2003.1263280.

[15] E. Bacon, N. Folic, N. Ioannides, I. Schagaev, “Curriculum design, development and assessment for computer science and similar disciplines,” FECS'12 - The 2012 Int’l Conference on Frontiers in Education: Computer Science and Computer Engineering, Las Vegas, July 2012, Paper ID: FEC4130.

[16] E. Bacon, N. Folic, N. Ioannides, I. Schagaev, “Multiple choice answers approach: Assessment with penalty function for computer science and similar disciplines,” International Journal of Engineering Education Vol.28, No.6, pp. 1-7, 2012.

[17] E. Bacon, G. Hagel, N. Folic, M. Charnine, R. Foggie, B. Kirk, I. Schagaev, G. Kravtsov, “Web-enhanced design of university curricula,” FECS'13 - The 2013 Int’l Conference on Frontiers in Education: Computer Science and Computer Engineering, Las Vegas, July 2013.