Personalized Subject Learning Based on Topic Detection and Canonical Correlation Analysis

Zhangzu SHI, Steve K. SHI, Lucy L. SHI
Smart Education Center, National Research Institute of Smart City and Big Data,
North 4th Ring Middle Road, Beijing, China

***Abstract*—To keep pace with the time, learning from printed medium alone is no longer a comprehensive approach. Fresh digital contents can definitely be the complement of printed education medium. Although timely access to fresh contents is becoming increasingly important for education and gaining such access is no longer a problem, the capacity for human teachers to assimilate such huge amounts of contents is limited. Topic Detection (TD) is then a promising research area that addresses speedy access of desired contents based on topic or subject. On the other hand, personalized education is getting more attention because it facilitates the improvement of creativity and subject learning of the students. This paper reveals a patented Personalized Subject Learning (PSL) system that caters for the need of personalized education and efficiently provides subject based contents. An efficient topic detection algorithm for providing subject content is presented. Moreover, since education contents are multimedia based ones with multimodal, PSL introduces Canonical Correlation Analysis (CCA) method to detect multimodal correlations across different types of media. Due to its novelty, PSL has been used as the key engine in a real world application of personalized education system as the smart education module sponsored by a Smart City project.**

***Keywords*—Topic Detection; Canonical Correlation Analysis; Personalized Education; Subject Learning; Multimodality**

I. INTRODUCTION

Throughout most of history, only the wealthy have been able to afford an education geared towards the individual learners. For the vast majority, education has remained a mass affair, with standard curricula, pedagogies, and assessments. It has been believed so long as the system insists on teaching all students the same subjects on printed medium in the same way, progress will be incremental. However, now for the first time it is possible to individualize education -- to teach each person what he or she needs and wants to know in ways that are most comfortable and most efficient, which may produce a qualitative spurt in educational effectiveness. How can we improve the performance in education, while cutting costs at the same time? In 1984, it was shown that individualized tutoring had a huge advantage over standard lecture environments: students who received individualized tutoring t performed better than 98 per cent of students from the standard classes. Yet the question is how to make individualized or personalized education affordable. Daphne Koller from Stanford AI Lab [1] argued that technology may provide a path to this goal.

Today timely access to fresh contents is becoming increasingly important in today's education, and gaining such access is no longer a problem because of the widespread availability of broadband both in homes and businesses. Ironically, high-speed connectivity and the explosion in terms of the volume of digitized textual content available have given rise to a new problem, namely, information overload. Clearly, the capacity for human teachers to assimilate such vast amounts of contents is limited. Topic Detection (TD) has emerged as a promising research area that harnesses the power of modern computing to address this new problem by helping us obtain desired subjects or personalized topics in an automatic way. A topic is defined as a seminal event or content, along with all directly related events and contents . Thus, it is inferred that a topic consists of events and contents, both of which are defined in greater detail [2]. A Topic Detection and Tracking (TDT) is defined as something that happens at a specific time and place, along with all the necessary preconditions and unavoidable consequences. Such an event might be a new movie, an election, or an alien attack. TD enables the automatic discovery of new topics from a news corpus and the subsequent assignment of news documents to the discovered topics [3]. A new topic typically corresponds to a newsworthy incident such as the 2012 US presidential election. Therefore, TD technology is a perfect tool for clustering fresh subject contents.

Moreover, education contents are usually multimedia based. They can be texts, animations, sounds, videos, and so on. Text based TD solution alone is not able to do the final content fusion for personalized contents recommendation. In the process of the cross-media recommendation, the query examples and recommended results need not to be of the same media types. For example, students can receive sound pieces by submitting either an image example or a sound example. This is the so called multi-modal environment. Canonical Correlation Analysis (CCA) is then accommodated to calculate the correlations and measure multi-modality similarities across media types [4].

Despite the fact that existing TD solutions play important roles in their applications [3, 5, 6, 7, 8, 9, 10, 11], they do not explicitly incorporate Language Model and cross-media CCA model into their formulations. Based on previous research [12, 13, 14, 15, 16, 17, 18], a novel personalized subject learning (PSL) system is created based on the above ideas. PSL system is a computer aided education system using TD technologies and CCA methodology. Enlightened by achievements in Information Retrieval (IR) field, Relevance Model (RM) is adopted as the language model for TD. RM is a theoretical extension of statistical language modelling and applicable in

both retrieval and TD [19]. By treating education contents as news and stories, both TD and IR methods can be used to retrieve relevant contents and feed them into CCA to analyse cross-media correlations.

The remainder of this paper is organized as follows: In Section 2, key concepts and terms are defined and works directly related to PSL system are reviewed. Section 3 describes a novel approach in terms of TD and CCA. In Section 4, the superiority of the approach of PSL is demonstrated. Finally, in Section 5, conclusions and some future research directions are presented.

II. Diagram and Example of Personalized Subject Learning

Personalized education refers to providing learning experiences tailored to each student's interests and learning styles. It also implies student-directed and self-managed learning. Teachers may individualize instruction in a classroom setting but admit that this is hard to accomplish given the competing need to cover subject matter material. Well programmed computers, whether in the form of personal computers or hand-held devices, are becoming an alternative choice. They will offer many ways to master materials. Students (or their teachers, parents, or coaches) will choose the optimal ways of presenting the materials. Appropriate tools for assessment will be implemented too. Most importantly, computers are infinitely patient and flexible. With a computer aided personalized subject learning system, human beings can spend the precious classroom time on more interactive problem-solving activities, which may help them achieve better understanding and foster creativity. Once the personalized education takes hold, the world will be very different. Many more individuals will receive better education because they will be learning knowledge in ways that suit them best.

A personalized subject learning system consists of 17 components, as shown in Figure 1. Components 1-4 capture various sorts of input from students, including motion, speech, drawing and text input. Component 7 performs efficient topic detection task and contents are fed in from various sources (e.g. Component 11-offline contents such as digital library, Component 12-contents edited by teachers (component 17), Component 13-online contents (learning from books alone is no longer the way to keep pace with the time). Fresh online contents are definitely the complement of printed education medium. Component 9 records learning behaviour of students and stores them into behavioural log. Contents that deserve to be education materials are collected by Component 10. The Component 6 analyses the correlations among collected multimedia contents and recommends the personalized subjects and contents to students and teachers. Component 8 is responsible for relevance feedback based on likes and dislikes of students and teachers as a mean to justify and improve the effectiveness of PSL.

The inputs and outputs of PSL system follow the sequence of Figure 1 To further elaborate theFigure, real examples are shown here in Figure 2 and Figure 3. Topics are detected and clustered as indicated by red arrows as shown in Figure 2 and then CCA decides which topic suits the personal needs of students as shown in Figure 3. In Figure 2, topics with "Old Summer Palace" have been detected and related contents are clustered to feed in the PSL system.

Precious antiques (textual and image contents) from Old Summer Palace have been returned to China as shown in the 1st and the 2nd pictures in Figure 3. The Film "12 Chinese Zodiac" directed by Jackie Chan (video content shown in the 3rd picture in Figure 3) has been co-related as teaching contents by CCA.

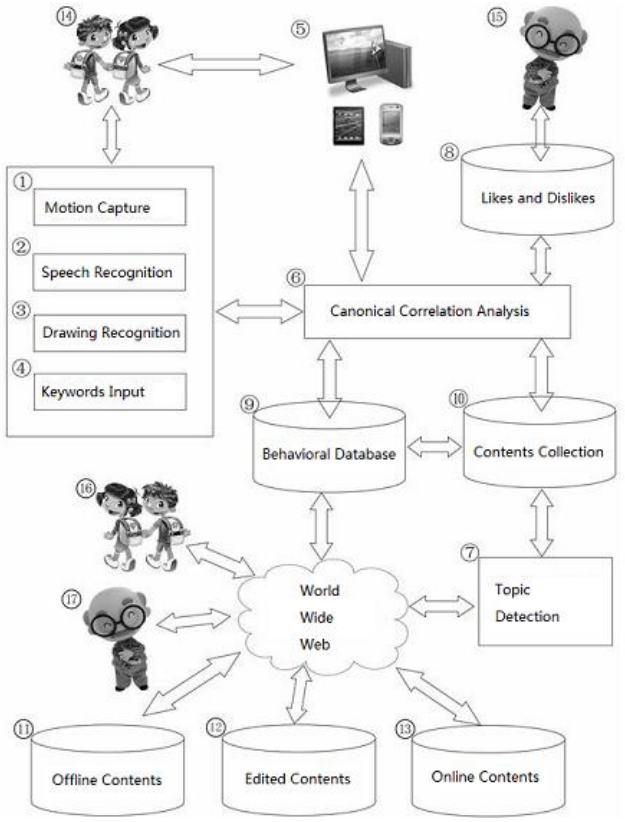

Fig. 1. Framework of Personalized Subject Learning System

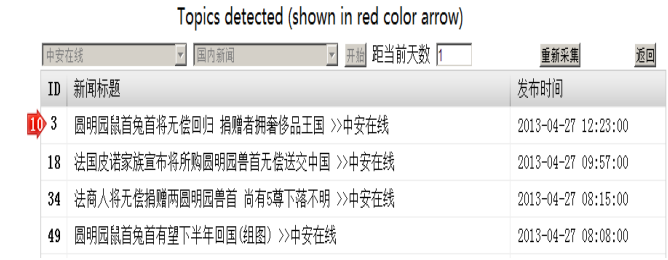

Fig. 2. Topics detected and clustered that are indicated by the red arrow

The system aims to provide personalized education materials based on subjects or topics. Traditional information retrieval (IR) system is not able to meet such a demand. Hence, this paper proposes a PSL system by accommodating efficient topic detection method and canonical correlation analysis method. The former shoulders the task of fast clustering documents from vast and multiple textual content sources into clustered subjects or topics. The latter is responsible for recommending relevant or co-related contents

Fig. 3. CCA Correlates Antiques Returned to Old Summer Palace and Films by Jackie Chan

by inter-media correlation measure and relevance feedback within the detected topics. Two methods work together to complement to each other for comprehensive, personalized and subject based interactions. Students and teachers then have the easy access to the vast amounts of personalized education contents anytime.

The following section of the paper illustrates the formal representation of two key components, that is, Component 7 for TD and Component 6 for CCA.

III. Formal Representations

In this section, the formal representations of PSL system especially for TD and CCA are described in order. Although there are many language tracking and modelling methods based on machine learning, thus far, the Vector Space Model (VSM) [20] has achieved the best results [21]. VSM has been successfully applied to the well-known SMART text retrieval system [22]. There are a number of formal ways of describing relevance feedback, beginning with the notion of an "optimal query" used in the SMART system. The biggest advantage of VSM is to simplify the text as the vector representation by its features and weights.

A. Document Representation

Contents of the document are expressed by a number of feature items, which generally include the basic linguistic units, such as words or phrases. $Document = D(t_1, t_2, ..., t_n)$, here t_k is a feature item. In a document, each feature item is assigned a weight w_k which denotes the feature item's degree of importance in the document:

$$D = D(t_1, w_1; t_2, w_2; ...; t_n, w_n) \quad (1)$$

Here the weight of t_k is w_k, and $1 \le k \le n$. Given a document $D = D(t_1, w_1; t_2, w_2; ...; t_n, w_n)$, a document can be expressed as a vector of n dimensional vector space. Expression $D = D(w_1, w_2, ..., w_n)$ is called as the Vector Space Model of D. The classic weight calculation method is $TF \times IDF$ in statistical methods. There are many ways to evaluate the significance of a term, ranging from simply identifying its existence to evaluating its distribution level in a document or in a whole corpus. The most common term weighting scheme for processing index terms is $TF \times IDF$, which stands for term frequency — inverse document frequency [21]. $TF \times IDF$ uses the term frequency and inverse document frequency of each feature item to calculate the weight. If tf_{ik} (Term Frequency) represents the number of occurrences of t_k in document D_i, idf_k donates inverse document frequency of t_k, then $TF \times IDF$ is defined as:

$$W_{ik} = tf_{ik} \cdot idf_k \quad (2)$$

Here tf_{ik} is a local statistic value which has different values in different documents. idf_k is a global statistic value reflecting a given term's distribution in all data set. The original definition of IDF is as follows:

$$idf_k = \log\left(\frac{N}{n_k}\right) \quad (3)$$

Here N represents the number of documents in all data sets, n_k represents the number of t_k that appears in data set. It can be seen that, the larger idf_k value is, the less the documents which contain the given term. If all documents contain the same given item, idf_k will be 0. In practice, to avoid such a case, equation (3) is improved by equation (4).

$$idf_k = \log\left(\frac{N}{n_k} + constant\right) \quad (4)$$

Generally, constant value is between 0 and 1, the equation (5) is then induced as:

$$idf_k = \log\left(\frac{N}{n_k} + 0.01\right) \quad (5)$$

If the document length on the impact of weights is taken into account, the feature item weights are normalized into the range of [0, 1]:

$$W_{ik} = \frac{tf_{ik} \times \log\left(\frac{N}{n_k} + 0.01\right)}{\sqrt{\sum_{k=1}^{n}\left[(tf_{ik}) \times \log\left(\frac{N}{n_k} + 0.01\right)\right]^2}} \quad (6)$$

B. TD Representation

The process of topic detection under this model is described here:

1) Topic is defined as $\vec{T} = (f_{T1}, f_{T2}, ..., f_{Tn})$, here $f_{Tj}(1 \le j \le n)$ represents the feature of topic T;

2) Follow-up story is defined as $\vec{d} = (f_{d1}, f_{d2}, ..., f_{dm})$, and here $f_{di}(1 \le i \le m)$

3) Represents the feature of news story $\vec{d}$;

4) Feature Selection is done by the following two steps:

- Stop words are removed;
- According to descending order of word frequency, the former *i* words are taken as feature items.

5) In TD research field, National Institute of Standards and Technology (NIST) and several universities, including Carnegie Mellon University (CMU), have been established benchmarks and corpus for TDT. In this paper, the similarity between $\vec{T}$ and $\vec{d}$ is defined as follows by adopting the principle reported by Lo and Gauvain of NIST [23]:

$$S(\vec{d},\vec{T}) = \frac{1}{L_d}\sum_{w\in d} tf(w,\vec{d}) \log \frac{\lambda P(w|\vec{T}) + (1-\lambda)P(w)}{P(w)} \tag{7}$$

Here $S(\vec{d},\vec{T})$ is the similarity of $\vec{T}$ and $\vec{d}$. w is the feature item of $\vec{T}$ and $\vec{d}$. $tf(w,\vec{d})$ is the frequency of w in $\vec{d}$. L_d is the whole number of terms in $\vec{d}$. λ is a smooth factor (0, 1) tuned to make the system achieve minimum cost when tracking TDT3 corpus. TDT3 corpus is created by NIST specially to accommodate Chinese news and stories. The smoothing technique is introduced to prevent data sparsity in unigram modeling.

$P(w|\vec{T})$ is the probability of w in $\vec{T}$.

$$P(w|\vec{T}) = \frac{C(w,\vec{T})}{Nw(\vec{T})} \tag{8}$$

$C(w,\vec{T})$ is the number of w occurrence in $\vec{T}$, $Nw(\vec{T})$ is the whole number of terms in $\vec{T}$, and $P(w)$ is a priori probability of w which is the statistic value in the background corpus.

$$P(w) = \frac{C(w,background)}{N(background)} \tag{9}$$

Here $C(w,background)$ is the number of w occurrences in background corpus; and $N(background)$ is the whole number of terms in background corpus.

6) According to similarity measurement of NIST, topic detection is then described as the calculation of the similarity between the story and the topic. In other words, if $S(\vec{d},\vec{T}) > \theta$, then they are considered as relevant or on-topic, off-topic otherwise.

C. Model Design of TD

Kullback-Leibler divergence is used to compute Relative Entropy (RE) as relevance measure between topic models to compensate the semantic weakness with similar aim of [24].

$$D(M_1 \| M_2) = \sum_w P(w|M_1) \log \frac{P(w|M_1)}{P(w|M_2)} \tag{10}$$

M1 and *M2* are the topic models for topic *T1* and *T2* based on RM**.** The two topic models, *M1 and M2,* both contain the word w . The equation (10) shows whether the two topic models *M1 and M2* have semantic similarity. When value *D* is close to 0, the similarity of two models is high. In order to enhance the robustness of the model, the Clarity probability is introduced for this case when both two models have smaller dissimilarity but they are similar to background corpus [25]. Such a phenomenon is called noise in that it is not a valid topic and therefore should be treated as a noise. Thus, equation (10) becomes the following one:

$$S(M_1 \| M_2) = \sum_w P(w|M_1) \log \frac{P(w|M_2)}{P(w|GE)} \tag{11}$$

Equation (12) is used in the experiment for more convenience of code design and equation (12) is a conversion of equation (11):

$$D(M_1 \| M_2) = \sum_w |P(M_1) - P(M_2)| + (1 - |P(M1) - P(GE)|) \tag{12}$$

Such a TD model design facilitates code design that then achieves linear performance with the combination of full text retrieval and new algorithm as shown in [16]. Other TD algorithms reported in literature have non-linear performance. The following experiments show lower error rates than those reported in [2].

D. CCA Representation

Content-based multimedia retrieval is a challenging issue, as it aims to provide an effective and efficient tool for searching media objects. Almost all of the existing multimedia retrieval techniques are focused on the retrieval research of single modality, such as image retrieval [26, 27], audio retrieval [28], video retrieval [29] and motion retrieval [30]. However, interactions that enhance students' engagement with Information and Communication Technology (ICT) are multimodal and include gesture, touch, language and so on. Due to the multiple modality of contents, an approach to extend cross-media retrieval to a more generalized multi-modality environment with less manual effort in collecting labeled sample data is needed. In this article, multi-modality representation [31, 32, 33] is adopted as it needs less manual effort in labeling multimedia documents already detected by TD module. In this subsection, the significance appears in inter-media correlation and solution of the problem of heterogeneous topics across different types of medium.

Co-relation of feature space X and feature space Y is defined as follows: $X_{(n\times p)}$ is denoted for n samples and p variables. $Y_{(n\times q)}$ is denoted for n samples and q variables. To obtain the main features, based on their feature weightage, a combination of variables from X and Y is extracted:

$$X_{(n\times p)} \xrightarrow{W_{x(p\times m)}} \mathrm{R}_{(n\times m)}; \qquad (13)$$
$$Y_{(n\times q)} \xrightarrow{W_{y(q\times m)}} \mathrm{S}_{(n\times m)}. \ (m < p \,\&\, m < q)$$

Here, W_x, W_y are subspace feature vectors. They are supposed to reduce the number of variables and use distribution of *R* and *S* to imitate that of *X* and *Y*. PSL uses relevance coefficient $\rho = r(\mathrm{R},\mathrm{S})$ as in (14) and is optimized by (15).

$$\rho = r(\mathrm{R},\mathrm{S}) = \frac{W_x{}^T C_{xy} W_y}{\sqrt{W_x{}^T C_{xx} W_x W_y{}^T C_{yy} W_y}} \qquad (14)$$

is the *covariance* matrix of $X_{(n\times p)}$ and $Y_{(n\times q)}$. Then with Lagrange multiplier method, $C_{xy}C_{yy}^{-1}C_{yx}W_x = \lambda^2 C_{xx}W_x$ is computed, which is a generalized Eigenproblem of the form $Ax = \lambda Bx$, and the sequence of W_x's and W_y's can be obtained by solving the generalized eigenvectors. Based on (13), minimum $\mathrm{R}_{(n\times m)}, \mathrm{S}_{(n\times m)}$ is computed to find out the correlation between $X_{(n\times p)}, Y_{(n\times q)}$. For example, let $x_i = (x_{i1},...,x_{ik},...,x_{ip})(x_{ik} \in \mathrm{Re}\,al)$ represents visual feature vector of motion (Component 1 of PSL) and $y_j = (y_{j1},...,y_{jk},...,y_{jq})(y_{jk} \in \mathrm{Re}\,al)$ represents feature vector of speech (Component 2 of PSL). Define x_i by subspace mapping as $x_i' = (x_{i1}',...,x_{ik}',...,x_{im}'), (x_{ik}' = a + b\times i, (a,b \in \mathrm{Re}\,al))$, y_j by subspace mapping as y_j'. Here, subspace is meant for Multi-modality Laplacian Eigen-Maps Semantic Subspace (MLESS).

Due to the existence of large quantity of complex numbers, coordinate values in each dimension of the subspace are converted to their polar form:

$$x_{ik}' = (\beta_{ik}, |x_{ik}'|) \qquad (15)$$
$$\beta_{ik} = arctg(b/a), |x_{ik}'| = \sqrt{a^2 + b^2}$$

The same conversion is done for y_j'. The semantic distance between motion x_i' and speech y_j' is then as follows:

$$CCAdis(x_i', y_j') = sqrt\sum_{k=1}^{m}(|x_{ik}'|^2 + |y_{jk}'|^2 \qquad (16)$$
$$-2\times|x_{ik}'|\times|y_{jk}'|\times Cos|\,\beta_{ik} - \beta_{jk}$$

PSL chooses the closest subject coupling with rich media contents and then provides recommends for students and teachers.

Topics are generally clusters of events and contents of specific subjects. To be personalized, clusters need to evolve as students and teachers learn more knowledge and the clusters are also able to optimize the feedback based on his or her experience, opinions, interests and creativity. In this way, personalized education material is finally achieved. In each evolution, the students or teachers have a chance to provide feedback regarding the recommended material and the feedback is treated as a guidance for next TD and CCA tasks.

IV. Experiments

A Java-based personalized education system [43] has been implemented. This system can be easily deployed on any Java virtual machine (JVM) platform.

A. Topic Detection

As a testbed, the system gathered news reports from standard testbed of NIST's TDT3 [2]. Besides, fresh rich media documents from Xinhua News Agency are also added. The experiments tested the viability of our work, in the context of real time fresh online and offline contents of NIST. Detection rate is justified by means of link detection task (LDT) as stated in [16].

B. Relevance Feedback of CCA

By adopting CCA approach co-researched with AI Lab of Zhejiang University, the experimental results of the relevance feedback of CCA [33] fully utilize the contents relevant to the detected topics or subjects, in the context of the user's opinions, creativity, personal knowledge and interests.

C. Practical Deployment

Practical deployment of our algorithm in real world is a patented system in both English and Chinese for personalized education as the smart education module of a Smart City project, as shown in Figure 4. Children's interactions with the computer were frequently referred to, by adult teachers and children, as "playing with the computer" in the same way as they would talk about playing with the bricks or the model animals. The personalized subjects are presented in front of the children as shown in lower portion of Figure 5. This is not surprising inasmuch as the dominant ethos of personalized environments is that children learn through play like a game format [34]: "Children's encounters with books, crayons, and paints were not referred to as play activities, probably because their role in the curriculum was easily identified and practitioners were used to recording children's development in the areas of reading, writing, and drawing. Children's freedom to choose resulted in highly varied patterns of engagement". With same opinions, three categories of teacher involvement, in PSL's computer play, are reactive supervision, guided interaction and a hybrid approach that combines the elements of both.

The application of PSL research investigated learning in personalized settings and an adapted version of the framework and fundamental technology breakthrough have the potential to become research tools and to support changes in practice for professionals in other sectors of education. For example, it is by no means a novel observation that families play a key role in supporting children's learning.

Published during the 1960s, the influential Plowden Report [35] has a section on the importance of parental attitudes and the 'physical amenities' at home. It is recognized that children acquire almost as much general knowledge at the home as in the school, and almost as much information about the world

and the way it works during leisure hours as from the formal lessons in the classroom.

Fig. 4. Personalized multimodal subjects are shown for the students and teachers in the implemented PSL system (smart education module of Smart City project)

Parents can play the role of teachers in PSL since there has been a clear extension in the trends of education from formal settings to the home and more parental engagement [34].

V. Conclusions and Further Works

Due to its efficiency and effectiveness, such a breakthrough meets the practical demands in the fields of Community Question Answering (CQA) [36], social link management [37, 38], learning for personal environment or R&D activities [39, 40, 41], preschool cognitive growth and hence, a distinguished patent has been granted [17].

Think about the guided interaction that helps practitioners to question the purpose of information and communication technology (ICT) and to articulate, reflect on and legitimize the changes in pedagogies. PSL prompts changes in the provision of resources, planning and assessment. Practitioners become more innovative, expand their definition of ICT as well as using existing resources in different ways, and begin to plan for, observe and record student's engagement with ICT in new ways. The breakthrough of PSL in this paper appears not only in the fast TD based clustering but also for CCA based measurable rich media topic recommendation towards subject learning with persistence, engagement and pleasure. Personalized Subject Learning is becoming the trend for people to learn fresh contents. This research shows the capacity and efficiency to automatically deal with vast amounts of information and contents. Hence, the PSL system shows obvious applicability and availability.

In the course of this work, a number of interesting questions have been encountered that we hope to answer in future research. Besides satisfying multimedia contents, the PSL system is able to process multilingual contents in one shot. The research team is currently working on 52 other languages besides English and Chinese. An international PSL system across countries should cater for such a need in the future. It is planned to have in depth collaboration with teams in the States, Europe and Singapore which are keen on PSL and aim to form an international personalized education alliance along with this endeavor.

Acknowledgment

This research is supported by the Chinese National Natural Science Foundation under the grant number of 61073150. The Smart Education Center played a guiding role during the engagement of the Personalized Education Program. The Personalized Education System is sponsored by the National Incubating Center of China and JoinVC Holdings as part of the EASE (Easy Active Systematic Education - 易智童) project, which was initiated under the National Strategy of Smart City [42]. To complement the System, an educational book named "英才是怎样练成的?" is in press, which is supported by JoinVC Holdings, Hangzhou JingWuMen Education Technology and Zhejiang University. The Artificial Intelligence Laboratory of Zhejiang University and Stanford provided us some constructive suggestions in CCA modeling. We would like to express our sincere thanks to all of them.

References

[1] KOLLER, D. , "Technology as a passport to personalized education", page D8 of the *New York Times*, December 6, 2011.

[2] TDT (2004).Topic Detection and Tracking. Annotation manual Version 1.2. http://www.nist.gov/speech/tests/tdt.

[3] BUN, K. K. and ISHIZUKA, M. "Topic extraction from news archive using TF*PDF algorithm," *Proc. of Third Int'l Conf. Web Information Systems Eng. (WISE '02)*, 2002, pp. 73-82.

[4] HARDOON, D. R., SZEDMAK, S. and SHAWETAYLOR, J.. "Canonical correlation analysis; An overview with application to learning methods," *Technical Report CSD-TR-03-02, Computer Science Department, University of London*.

[5] YANG, Y, PIERCE, T. and CARBONELL, J. ."A Study of retrospective and on-line event detection," *Proc. of the 21st annual international ACM SIGIR conference on Research and development in information retrieval, ACM SIGIR '98*, 1998, pp. 28-36.

[6] ALLAN, J., LAVRENKO, V. and JIN, H.. "First story detection in TDT is hard," *Proc. of Ninth Int'l Conf. Information and Knowledge Management*, 2000, pp. 374-381.

[7] STOKES, N. and CARTHY, J. " Combining semantic and syntactic document classifiers to improve first story detection," *Proc. of the 24th annual international ACM SIGIR conference on Research and development in information retrieval, ACM SIGIR '01*, 2001, pp. 424-425

[8] BRANTS, T., CHEN, F. and FARAHAT, A"A system for new event detection," *Proc. of the 26th annual international ACM SIGIR conference on Research and development in information retrieval, ACM SIGIR '03*, 2003, pp. 330-337.

[9] KUMARAN, G. and ALLAN, J. "Text classification and named entities for new event detection," *Proc. of the 27th annual international ACM SIGIR conference on Research and development in information retrieval, ACM SIGIR '04*, 2004, pp. 297-304.

[10] CHEN, K.-Y., LUESUKPRASERT, L. and CHOU, S. T. "Hot topic extraction based on timeline analysis and multidimensional sentence modelling," *IEEE Transactions on Knowledge and Data Engineering*, **19**(8), 2007, pp. 1016-1025.

[11] HE, Q., CHANG, K., LIM, E.-P. and BANERJEE A. "Keep it simple with time: A re-examination of probabilistic topic detection models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **32**(10), 2010, pp. 1795 – 1808.

[12] SHI, K., HE, J., LIU, H., ZHANG, N. and SONG, W. "Efficient text classification method based on improved term reduction and term weighting," *The Journal of China Universities of Posts and Telecommunications*, Vol **18**, 2011, pp. 131-135.

[13] SHI, K., LI, L., HE, J., LIU, H., ZHANG, N. and SONG, W. "A linguistic feature based K-means text clustering method," *Proc. of IEEE Cloud Computing and Intelligent Systems*, 2011, pp. 108-112.

[14] SHI, K., LI, L., HE, J., ZHANG, N., LIU, H. and SONG, W. "Improved GA-based document clustering algorithm," *Proc. of IEEE Broadband and Multimedia Communications*, 2011, pp. 675-679.

[15] SHI, K., LI, L., LIU, H., HE, J., ZHANG, N. and SONG, W. "An improved KNN text classification algorithm based on density," *Proc. of IEEE Cloud Computing and Intelligent Systems*, 2011, pp. 113-117.

[16] SHI, K. and LI, L. "A Close-to-linear Topic Detection Algorithm using Relative Entropy based Relevance Model and Inverted Indices Retrieval," *International Journal of Computational Intelligence Systems*, **5**(4), 2012, pp. 735-744.

[17] SHI, K. and SHI, Z. "Subject Shifting based on the Consciousness and Current Focus of Audiences," Patent, 2012.

[18] SHI, K. and LI, L. "High performance genetic algorithm based text clustering using parts of speech and outlier elimination," *International Journal of Applied Intelligence*, Vol 38, Issue 4, 2013, pp 511-519..

[19] CROFT, W. B., CRONEN-TOWNSEND, S. and LAVRENKO, V. "Relevance feedback and personalization: A language modelling perspective," *Proc. of the DELOS-NSF Workshop on Personalization and Recommender Systems in Digital Libraries*, 2001, pp. 49-54.

[20] SALTON, G., WONG, A. and YANG, C.S."A vector space model for information retrieval," *Communications of the ACM*, **18**(11), 1975, pp. 613–620.

[21] SALTON, G. and YANG, C.S. "On the specification of term values in automatic indexing," *Journal of Documentation*, 29(4), 1973, pp. 351-372

[22] SALTON, G. *Automatic Information Organization and Retrieval*. New York, 1968, NY: McGraw-Hill.

[23] LO, Y. and GAUVAIN, J. "The LIMSI Topic Tracking System for TDT2001," *Topic Detection and Tracking Workshop*, Gaithersburg, MD, National Institute of Standards and Technology, 2001.

[24] LEE, C., LEE, G. G. and JANG, M. "Dependency structure language model for topic detection and tracking," *Information Processing and Management***43**(5), 2007, pp. 1249–1259.

[25] LAVRENKO, V., ALLAN, J. and DeGuzman, E. "Relevance models for topic detection and tracking," *Proc. of the Human Language Technology Conference*, 2002, pp.104–110.

[26] CHANG, E., GOH, K., SYCHAY, G. and WU, G. "CBSA: Content-based soft annotation for multimodal image retrieval using Bayes point machine," *IEEE Transactions on Circuits and Systems for Video Technology*, **13**(1), 2003, pp. 26-38.

[27] HE, X., MA, W. Y. and Zhang, H. J. "Learning an Image Manifold for Retrieval," *Proc. of the 12th annual ACM international conference on Multimedia*, 2004, pp. 17-23.

[28] GUO,G. and LI, S.Z. "Content-based audio classification and retrieval by support vector machines," *IEEE Transactions on Neural Networks*, **14**(1),2003, pp.209-215.

[29] FAN, J., ELMAGARMID, A. K., ZHU, X., AREF, W. G. and WU, L. "ClassView: hierarchical video shot classification, indexing, and accessing," *IEEE Transactions on Multimedia*, **6**(1), 2004, pp. 70-86.

[30] MÜLLER, M., RÖDER, T. and CLAUSEN, M. "Efficient content-based retrieval of motion capture data," *ACM Transactions on Graphics*, **24**(3), 2005, pp. 677-685.

[31] WU, F., YANG, Y., Zhuang, Y. and Pan, Y. "Understanding multimedia document semantics for cross-media retrieval," *Proc. Of the 6th Pacific-Rim conference on Advances in Multimedia Information Processing - Volume Part I, PCM'05*, Berlin, Heidelberg: Springer-Verlag, 2005, pp. 993-1004.

[32] ZHANG, H. and WENG, J. "Measuring multi-modality similarities via subspace learning for cross-media retrieval," *Proc. of the 7th Pacific Rim conference on Advances in Multimedia Information Processing,PCM'06*, Berlin, Heidelberg: Springer-Verlag, 2006, pp. 979-988.

[33] ZHUANG, Y., WU, F., ZHANG, H. and YANG, Y. "Cross-Media Retrieval: Concepts, Advances and Challenges," *Proc. of 2006 International Symposium on Artificial Intelligence*. Vol 4261, the series Lecture Notes in Computer Science, 2006, pp 979-988.

[34] Lydia Plowman and Christine Stephen, "Children and computers in pre-school", *British Journal of Educational Technology*, Vol 36 No 2, 2005 , pp.145 – 157.

[35] CACE - The Central Advisory Council for Education, "Children and their Primary Schools: A Report of the Central Advisory Council for Education (England)", 1967.

[36] ZHANG, Z. and LI, Q. "Hot topic discovery and trend analysis in community question answering systems," *Expert Systems with Applications*, **38**(6), 2011, pp. 6848–6855.

[37] GARCÍA-CRESPO, A., COLOMO-PALACIOS, R., GÓMEZ-BERBÍS, J. M. and GARCÍA-SÁNCHEZ, F. "SOLAR: Social link advanced recommendation system," *Future Generation Computer Systems*, **26**(3), 2010, pp. 374-380.

[38] GARCÍA-CRESPO, A., COLOMO-PALACIOS, R., GÓMEZ-BERBÍS, J. M. and RUIZ-MEZCUA, B. "SEMO: A framework for customer social networks analysis based on semantics," *Journal of Information Technology*, **25**(2), 2010, pp. 178-188.

[39] COLOMO-PALACIOS, R., GARCÍA-CRESPO, Á., SOTO-ACOSTA, P., RUANO-MAYORAL, M. and JIMÉNEZ-LÓPEZ, D. "A case analysis of semantic technologies for R&D intermediation information management," *International Journal of Information Management*.**30**(5), 2010, pp.465-469.

[40] GARCÍA-PEÑALVO, F. J., CONDE-GONZÁLEZ, M. Á., ALIER-FORMENT, M. andCASANY-GUERRERO, Mª J. "Opening Learning Management Systems to Personal Learning Environments," *Journal of. Universal Computer Science,***17**(9), 2011, pp.1222-1240.

[41] GARCÍA-PEÑALVO, F. J., ORDÓNEZ DE PABLOS, P., GARCÍA, J. and THERÓN, R. "Using OWL-VisMod through a decision-making process for reusing OWL ontologies," *Behaviour & Information Technology*. Vol 33, Issue 5, 2014, pp. 426-442.

[42] SHI, Z. and SHI, K., 易智童 - Personalized Education System, http://www.joyscan.com, www.joypond.com, www.joinvc.com, 2015.

[43] SHI, Z. and SHI, K., "英才是怎么炼成的"- How the Elite was Tempered, book in press, December 2015.