A New Intelligent Topic Extraction Model on Web

Ming Xie
Computer School of Wuhan University, Wuhan, China; Guangxi Economic Management Cadre College, Nanning, China
Email: wolfgangtse@gmail.com

Chanle Wu* and Yunlu Zhang
Computer School of Wuhan University, Wuhan, China
Email: wuchle@whu.edu.cn, zhang.yunlu850527@gmail.com

Abstract—**We tackle the problem of topic extraction on Web. In this paper, we propose an approach to implementing ontology-based data access in WordNet with the distinguishing feature of optimizing density-based clustering OPTICS algorithm (DBCO) to extract topics. Our solution has the following two desirable properties: i) it uses WordNet for word sense disambiguation of words in the learning resources documents and ii) it mapping the data space of the original method to a vector space of sentence, improving the original OPTICS algorithm. We outline the interface between our scheme and the current data Web, and show that, in contrast to the existing approaches, no exponential blowup is produced by the DBCO. Based on the experiments with a number of real-world data sets of 310 users in three study sites, we demonstrate that topic extraction in the proposed approach is efficient, especially for large-scale web learning resources. According to the user ratings data of four learning sites in the 150 days, the average rate of increase of user rating after the system is used reaches 25.18%.**

Index Terms—**Topic Extraction, E-learning, Semantic, Ontology**

I. INTRODUCTION

With the rapid development of the Internet, there are more and more course information published, which brings in the coexistence problems of "information overload" and "knowledge poor". It costs much time and manpower to manually extracting topic information from unstructured text or multimedia with existing methods. How to implement topic auto-extract in unstructured learning resources, so that users quickly and accurately obtain the knowledge they want, is the key issues that the next generation e-learning theory and technology should focus on. The paper discusses the existing topic auto-extraction technologies, through which, large amount of course information is presented to user concisely and accurately. In order to eliminate redundant information in curriculum, this paper proposes a clustering algorithm-based topic auto-extraction knowledge model. This model applies information fusion technology into the extraction process of the contents of topics. This paper designs an improved OPTICS algorithm (NOP)-based Multi-document automatic summarization system, which contributes in two aspects: (1) improving the sentence similarity computing method, computing sentences semantic similarity according to the semantic relations between among words in sentence, and based on this clustering sentence, complete the division of the sub-themes, and finally extract a number of sentences in various sub-themes as a summary sentence with certain strategy, extract a certain number of sentences as topic description sentence. colleting the theme concept rather than the word form, using semantic resources WordNet for word sense disambiguation of words in the learning resources documents, and then extracting the theme concept to build vector space model for topic extraction; (2) improving the original OPTICS algorithm, mapping the data space of the original method to a vector space of sentence, and gives an approach of repositioning sparse nodes. Proposing a density-based clustering algorithm, applying it into web course systems; It divides multi-document collections into different sentence clusters; then extracts a certain number of sentences from different sub-themes to produce the digest.

Definition1. Given a specified norm topic ***T*** is defined as a triple: ***T={T_id, T_contenet, T_tag}.***

For example, we can extract the ***T*** of “software” from the segment of document such as “*Software is a set of items or objects that form a “configuration” that includes programs, documents and data*”. According to that, the ***T(software)= {T_id(software), T_contenet(software), T_tag(software)}.*** *The* ***T_id(software)*** is given by the system; the ***T_contenet(software)*** is “*a set of items or objects that form a configuration” that includes programs, documents and data*”, that will be extracted from the document; ***T_tag(software)*** is “Software=program+ document”.

Definition2. **Intelligent Topic Extraction on Web** is defined as a process to get the content of topic from the web resource automatically.

For example, we can use the proposed model to extract

© 2011 ACADEMY PUBLISHER
doi:10.4304/jcp.6.3.466-473

the content of topic from the primitive documents on web and store it in the triple: ***T={T_id, T_contenet, T_tag}.***

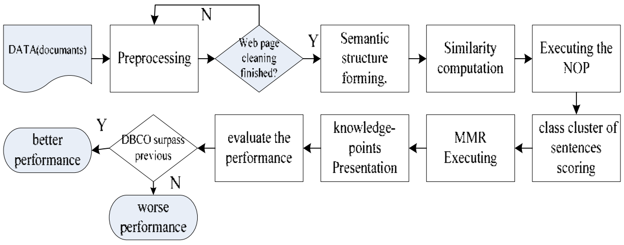

Figure 1. The scheme of topic extraction

II. Related work

Kim et al. used a special word sense disambiguation technique [1]. In this method, only the 25 most original meanings (Root sense) in WordNet are considered, each word is assigned with a meaning so as to ensure the accuracy of disambiguation. Although they did not give the accuracy of disambiguation, the experimental results on TREC7 and TREC8 data showed over 10% of the increase in the disambiguation performance. Their attempt to add semantic information in the BM25 formula also has been successful. Liu et al. tried to eliminate the ambiguity of the query from an opposite viewpoint [2]. The disambiguation experiment is based on 250 queries evaluated by TREC13Robust, the results show that, Liu’s disambiguation method can disambiguate all the 333 ambiguous words in queries, the accuracy was 90%, and the retrieve results is increased by 10% to 25% in five TREC data sets[3]. Michael Ankerst proposed the OPTICS algorithm with the characteristic of density-based clustering across data sets. This algorithm enables digging data of corpus structure [4]. Since the cosine similarity is more concerned about the consistency of the vector direction, which is the characteristic of consistency, the word feature vectors and the cosine similarity are more suitable for calculation of the sentence similarity. We can choose which features the word feature vectors and meaning to both the analysis and calculation, the original method and vocabulary words from the vector were replaced with the concept of distance vector and concepts. Newsblaster is a system developed by Columbia University in multi-document summarization [5], which is a news tracking tool to make daily major news-related abstracts. NeATS [6] is another system of multi-document summarization technology developed by University of Southern California. This system extracts important concepts through reliable statistical information, according to the information of the beginning word in sentences and location information of sentences. Because some sentences begin from a conjunction or verb phrase, if those sentences are extracted in abstract sentence, then the inconsistency will be weaken, so it is needed to filter out these sentences by using the MMR method and reasonable sort of the abstract sentence (such as in chronological order) in order to produce abstracts. Compared with the multi-document summarization technology, the extraction of topics put more requirements on the themes concept clustering. Research such as [7] targeted on specific content extracting, which integrated knowledge extraction and knowledge mining to extract knowledge from text, and then discover knowledge by association rules. Reference [8] determined directive words from the symbol level. Its extraction objects, however, must be structured and do not meet the requirements of the topic extraction. Paper [9] proposed a reference point and density-based fast clustering algorithm, but it is not appropriate to the dimension reduction of high-dimensional document vector space. Zeng et al. applied the OPTICS algorithm into text clustering [10]. Zhao proposed a similarity measure approach based on different characteristics of sentences [11]. However, the multi-theme documents are major part in multi-document collection, if the sentence is taken by traditional methods; it is likely to ignore certain information in the document collection. There are multiple documents in similar classes, in a multi-document collection, different authors descript the same knowledge in different angles sometimes, or even in the opposite angles, which makes more than one themes appears in multi-document collection. Documents in multi-document collection associate within each other through a common theme, which is taken as central theme. Sub-theme of multi-document collection is the sentences combination with same meaning. these sub-themes present various local information of the document collection. Therefore sub-theme division will be a specific issue in multi-document summarization.

III. The Framework of Intelligent Topic Extraction in Web Learning Resources

Aiming at providing solutions for the important problems of heterogeneous, discrete, "information overload" and "knowledge poverty" in Web learning resources, we propose a topic auto-extraction method and construct topic auto-extraction system for Web learning resources. By improving and optimizing density-based clustering OPTICS algorithm, we propose the NOP algorithm by mining the internal relations in document collection, divide the document theme more accurately, and improve the topic extraction results. This paper designs a sentence clustering-based topic auto-extraction system by improving the density-based OPTICS algorithm. The system cluster by computing the similarity of sentences, together sentences with the same theme so that each class is represented as a sub-theme document collection in learning resources, and then extract certain number of sentences from each sub-theme to generate description sentence of topic content, finally presents topic contents to the learners. The system's global architecture is shown in Fig. 1.

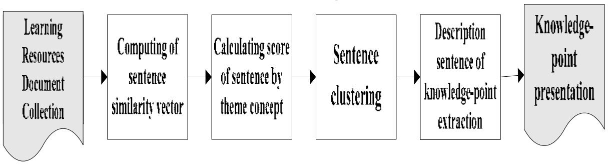
Learning Resources Document Collection
Computing of sentence similarity vector
Calculating score of sentence by theme concept
Sentence clustering
Description sentence of knowledge-point extraction
Knowledge-point presentation

Figure 2.The global structure of topic extraction

© 2011 ACADEMY PUBLISHER

Problem Definition

1. There is ambiguity in topic descriptions and topic information inputs of learning resources. To solve the problems of topic expression ambiguity by WordNet description of semantic structure, merging layer by layer, realize semantic disambiguation.

2. The OPTICS algorithm for the data space modeling is applied to the document space model, and the traditional word form-based similarity computing cannot discover the deep information among words. By analyze and calculate both the word characteristic vector and the word meaning vector, we can develop a new approach to replace the measure of word vector and words distance, which are used in the traditional extraction. We need to solve the problem of feasibility and accuracy of Topic extraction, and provide a semantic-level data mining method to avoid the high cost and dispersion of manual extraction.

The steps of topic extraction in learning resources are as following.

Step1.Preprocessing. For example, to accomplish web page cleaning in web learning resources, the resources should be stored as a text document.

Step2. Web page cleaning. The sentences of the of text documents are filtered.

Step3. Semantic structure forming. The semantic structures are described by WordNet, merged layer by layer, to complete semantic disambiguation.

Step4. Similarity computation. This work is done through the concept characteristic vector and concept distance vector among sentences.

Step5. Executing the NOP (an improved OPTIC algorithm) to clustering.

Step6. Scoring the class cluster of sentences.

Step7. Executing the MMR (Maximal Marginal Relevance) method, to choose the sentences of describing topic content, that have high relevance degree to themes while the redundancy among the sentence and other chosen ones is as small as possible.

Step8. Presentation of the content of topics.

Step9. Reference to Edmundson evaluation method, designing two Category of experiments, which evaluate the performance of topic extraction system from subjective, and objectively aspects. The global technology roadmap is shown in Fig. 2.

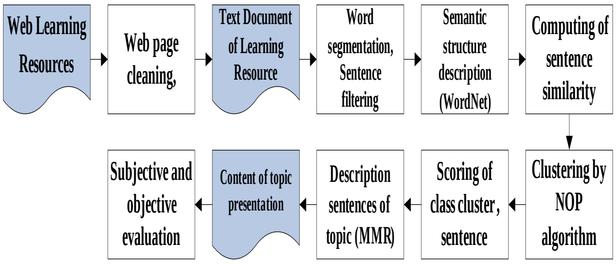
Web Learning Resources
Web page cleaning,
Text Document of Learning Resource
Word segmentation, Sentence filtering
Semantic structure description (WordNet)
Computing of sentence similarity
Clustering by NOP algorithm
Scoring of class cluster , sentence
Description sentences of topic (MMR)
Content of topic presentation
Subjective and objective evaluation

Figure 3. The global technology roadmap

IV. Algorithm and Computing

Notation:

$\boldsymbol{R_u}$ user's query*;* ***T*** topic*;* ***C( )*** web page cleaning; ***R*** web learning resources; ***D*** text document; ***F( )*** sentence filtering; ***S( )*** Semantic structure described by WordNet; ***M( )*** Merging layer by layer; ***Disa( )*** semantic disambiguation; $\boldsymbol{Sen_{CFT}}$ the sentences after the procedures of web page cleaning, sentence filtering and semantic disambiguation, which is stored as text documents; ***Sim( )*** similarity computation; $\boldsymbol{V_{cc}( )}$ concept characteristic vector; $\boldsymbol{V_{cd}( )}$ concept distance vector; ***NOP( )*** the NOP clustering(improved OPTIC algorithm); $\boldsymbol{Clu_c}$ class cluster; ***Scor( )*** Scoring; ***Rank( )*** Ranking; $\boldsymbol{Sen_R}$ the sentences after the procedures of Ranking; $\boldsymbol{Rel_{MMR}( )}$ relevance degree computed by the MMR (Maximal Marginal Relevance) method; $\boldsymbol{Red_{MMR}( )}$ redundancy degree computed by the MMR (Maximal Marginal Relevance) method; $\boldsymbol{T_{Rel}}$ threshold of relevance degree to themes; $\boldsymbol{T_{Red}}$ threshold of redundancy degree to themes; $\boldsymbol{Sen_d}$ the chosen sentences of describing the topic content; ***P( )*** the content; $\boldsymbol{Eval_s( )}$ a procedure to evaluate the performance of the topic extraction system from the subjective aspect; $\boldsymbol{Eval_o( )}$ a procedure to evaluate the performance of topic extraction system from the objective aspect.

Algorithm1. The Global Algorithm of Topic Extraction

```
Input: web learning resources $R$
Output: topic present $P(T)$
1. Begin
2. While(input(R)==True|| $PP!=False$){
3. $Sen_{CFT} \leftarrow Disa(M(S(F(C(R)))))$
4. $NOP \leftarrow Sim(V_{cc}(Sen_{CFT}), V_{cd}(Sen_{CFT}))$
5. $Clu_c \leftarrow NOP(Sim(V_{cc}(Sen_{CFT}), V_{cd}(Sen_{CFT})))$
6. $Sen_R \leftarrow Rank(top\ i,\ Scor(Clu_c),\ Scor(Sen_{CFT}))$, $1 \le i \le n$
7. $Sen_d \leftarrow \forall\ Sen_R\{Sen_R \mid Rel_{MMR}(Sen_R) \ge T_{Rel} \wedge Red_{MMR}(Sen_R) \le T_{Red}\}$
8. $P(T) \leftarrow P(Sen_d)$ }
9. End
```

Parameter:

C_i theme concept of a sentence; V vector space; V_1 vector of sentence S_1; ω_i the numbers that the theme concept Ci occurrences in S_1; V_2 vector of sentence S_2; φ_i the numbers that the theme concept C_i occurrences in S_2; ***similarity( )*** the similarity of word characteristics by word meaning distance; $\boldsymbol{Similarity_d( )}$ the concept of distance vector similarity of two sentences S_1 and S_2; $\{X_1, X_2, \ldots, X_i\}$, $(1 \le i \le m)$ the concept set of sentences S_1; $\{Y_1, Y_2, \ldots, Y_j\}$, $(1 \le j \le n)$ the concept set of sentences of S_2; $Distance(X_i, Y_j)$ the distance between the concepts X_i *and* Y_j; ***SIM( )*** sentence similarity computation; $\alpha = 0.7$ the coefficient of the concept characteristic vector; $\beta = 0.3$ that of the concept distance vector;

© 2011 ACADEMY PUBLISHER

Algorithm 2. Sentence similarity computing

```
Input: Sentence $S_1$, $S_2$
Output: SIM($S_1$, $S_2$ )   similarity of $S_1$ and $S_2$
1. Begin
2. $Similarity(S_1,S_2)=\overline{V_1}\cdot\overline{V_2}=\sum_{i=1}^{n}\omega_i\phi_i \Big/ \sqrt{\sum_{i=1}^{n}\omega_i^2} * \sqrt{\sum_{i=1}^{n}\phi_i^2}$
3. $\mathrm{Similarity}_d(S_1,S_2)=1\Big/\sum_{i=1}^{m}\sum_{j=1}^{n}Distance(X_i,Y_j)$
4. $SIM(S_1,S_2)=\alpha*Similarity(S_1,S_2)+\beta*Similarity_d(S_1,S_2)$
5. End
```

In the original DBCO algorithm, the ordered queue is always sorted in ascending order according to reachable distances. Therefore, the algorithm always selects to deal with the points with the smallest distance.

Parameter:

V a threshold value, when sentences in the class cluster is greater than or equal to V, the class cluster is taken as valid cluster; otherwise as invalid cluster. V is notated as $0.8*\mathrm{Avg}_{\mathrm{stNum}}(C)$; ***NUM$_v$(Sen)*** the number of valid sentences in C_i; ***NUM$_{vc}$(Sen)*** the number of valid sentences in all class clusters ***;*** ***C_i*** the ith sub-theme; ***Num$_i$(sen)*** the number of sentences in the class; ***Num$_i$(doc)*** the number of the original documents whose sentences belong to the sub-theme; ***Num(sen)*** the total number of sentences in the original document; ***Num(doc)*** the number of documents contained in the original document***;*** ***SCORE( )*** the class cluster score; ***Len( )*** length; ***Rank$_{rel}$( )*** ranking of the relevance degree of the sub-themes and the central theme***;*** ***Sen*** sentence***;*** ***C*** sub-theme; ***Sen$_d$*** the chosen sentences of describing topic content***;*** ***Len$_{inx}$(T)*** the minimum length requirements of topic description summery; and ***P( )*** the content.

A. Experiment

In this section we describe the experiments that we carried out on real-world web learning resources data sets. We chose these sets because they are publicly available. We take Edmundson as the evaluation method; design 2 types of experiments to evaluate the performance of our topic extraction system: (1) Subjective evaluation: In order to the subjective evaluation, we used the method of artificial extraction (by the domain experts and authorized teachers jointly) and the method of auto-extraction to extract the content of topics inspective in two courses"the interface technology of computer" and "the architecture of computer"(http://grid.whu.edu.cn/). The domain experts compare the extracted information with the two methods, and then give a review score. The rating levels are as follows: totally not similar, similar, very similar, completely similar to the other. (2) Objective evaluation: in the experimental evaluation, six tasks are tested, including DUC2005 and ROUGE 1.5.5 data sets, select ROUGE-N (where N take 1 to 4), ROUGE-L and ROUGE-W-1.2. There are voting systems in our topic extraction System, which are designed to collect the users' feedback. The users' review is presented in scores (0-100). If the user thinks the platform is really helpful for them in topic extraction, they evaluate it with a high score. We collect the users' remarks in 3 learning web sites in 150 days, and compare the average evaluation scores, before and after the system are used. The results are showed in Table 1.

In order to evaluate the performance of the topic extraction system, we record 8 experts' reviews on the topic extraction performance of 4 learning sites. The experiment result is showed in Table 2. As showed in Fig.4, different experts present their evaluations on the performance of our topic extraction system in the 3 learning sites. In general, the reviews on site 2 are better than those on other sites. However, the experts' reviews on site 3 are consistent with those on others.

B. Evaluation

As shown in Fig. 4,5,6,7,8,9,10,11, the scores of the 3 sites after using our system are better than those before using the proposed platform. The user ratings of the four learning sites are raised compared to the previous system before using the topic extraction system. The maximum increase is 36% in site 3; while the minimum increase is 17.91% in site 2; the average increase rate of user ratings after the system is used reaches to 25.18%. The result for the real-world data sets verifies the effectiveness of the proposed method. However, we notice that the improvement of the users' reviews in the experiment may be affected by other factors. For example, site 3 is the worst one in users' reviews before using the proposed approach. The suddenly increase of the scores might be partly from the contribution of the advertisement promotion during the days experiments are carried out. At least the results of 2 sites in our experiment clearly demonstrate that the proposed topic extraction system significantly outperforms the previous one. The results for the real-world data sets support this conclusion. All the dataset can be downloaded in the webpage http://grid.whu.edu.cn/rainbow/

We can see from Fig 12,13,14,15, there are some differences among the experts' reviews on the topic extraction performance, which illustrate that the stability of the topic extraction is still needed to be enforced. On the other hand, more objective work is needed in the future.

TABLE 1. THE AVERAGE EVALUATION SCORES OF 3 LEARNING SITES, (S1:SITE1, THRESHOLD VALUE *V=0.48*)

	S1	S2	S3
Pre	63	67	71
Cur	76	79	84

TABLE 2. THE AVERAGE EVALUATION SCORES OF 3 LEARNING SITES, (S1:SITE1, THRESHOLD VALUE *V=0.52*)

	S1	S2	S3
Pre	60	73	79
Cur	78	82	86

© 2011 ACADEMY PUBLISHER

Table 3. The average evaluation scores of 4 learning sites, (S1:SITE1, threshold value *V=0.56*)

	S1	S2	S3
Pre	76	74	82
Cur	84	87	90

Table 4. The average evaluation scores of 4 learning sites, (S1:SITE1, threshold value *V=0.6*)

	S1	S2	S3
Pre	72	79	67
Cur	81	87	84

Table 5. The average evaluation scores of 4 learning sites, (S1:SITE1, threshold value *V=0.64*)

	S1	S2	S3
Pre	75	78	70
Cur	80	83	82

Table 6. The average evaluation scores of 4 learning sites, (S1:SITE1, threshold value *V=0.68*)

	S1	S2	S3
Pre	68	71	74
Cur	75	78	80

Table 7. The average evaluation scores of 4 learning sites, (S1:SITE1, threshold value *V=0. 72*)

	S1	S2	S3
Pre	70	72	69
Cur	74	78	80

Table 8. The average evaluation scores of 4 learning sites, (S1:SITE1, threshold value *V=0.76*)

	S1	S2	S3
Pre	78	76	75
Cur	83	85	81

Table 9. Experts reviews on the topics extraction (S1:SITE1, E1:expert1, *V=0.48*)

	E1	E2	E3	E4	E5	E6	E7	E8
S1	76	74	68	73	72	71	67	68
S1'	82	76	71	79	77	80	75	81
S2	75	77	74	76	72	70	71	69
S2'	83	89	90	78	79	82	83	78
S3	74	73	72	75	73	72	68	70
S3'	84	86	85	88	80	81	79	76

Table 10. Experts reviews on the topics extraction (S1:SITE1, E1:expert1, *V=0.56*)

	E1	E2	E3	E4	E5	E6	E7	E8
S1	65	67	69	71	74	72	64	68
S1'	75	78	73	79	76	84	78	77
S2	67	70	71	72	68	66	69	73
S2'	78	82	83	77	75	80	81	77
S3	70	68	66	72	76	71	67	65
S3'	81	79	78	82	84	85	80	74

Table 11. Experts reviews on the topics extraction (S1:SITE1, E1:expert1, *V=0.64*)

	E1	E2	E3	E4	E5	E6	E7	E8
S1	72	74	67	70	73	68	69	71
S1'	78	76	71	77	79	81	80	76
S2	70	78	75	76	71	75	73	74
S2'	81	82	87	84	80	79	80	81
S3	76	80	74	73	77	75	74	72
S3'	84	82	83	78	80	82	84	76

Table 12. Experts reviews on the topics extraction (S1:SITE1, E1:expert1, *V=0.72*)

	E1	E2	E3	E4	E5	E6	E7	E8
S1	68	67	69	74	70	66	73	71
S1'	74	73	76	82	77	67	78	75
S2	72	74	78	75	71	73	76	77
S2'	80	78	83	82	85	80	77	79
S3	70	72	75	78	71	69	74	68
S3'	73	78	79	82	76	79	77	73

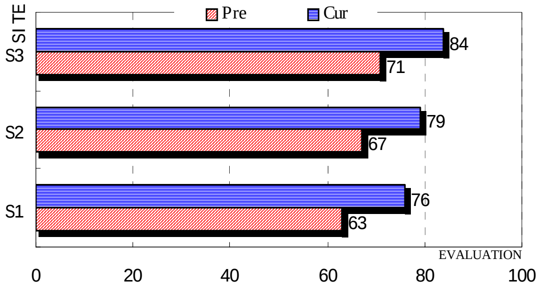

Figure4. Average evaluation scores of 3 learning sites (***V=0.48***)

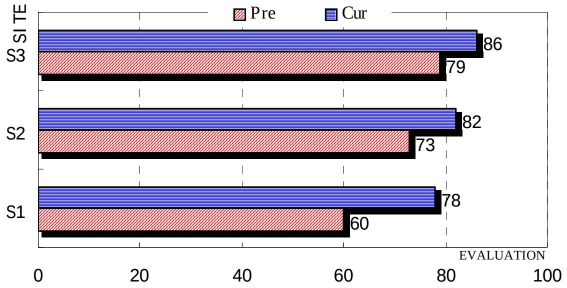

Figure5. Average evaluation scores of 3 learning sites (***V=0.52***)

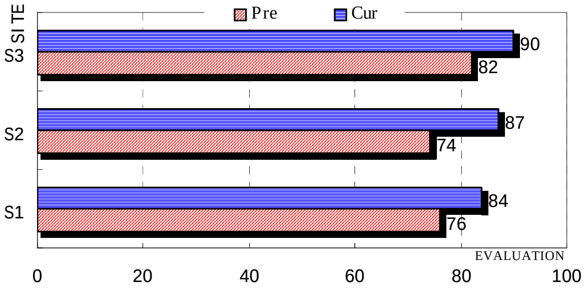

Figure6. Average evaluation scores of 3 learning sites (***V=0.56***)

© 2011 ACADEMY PUBLISHER

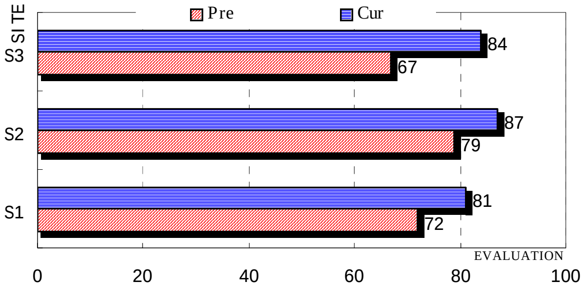

Figure7. Average evaluation scores of 3 learning sites (***V=0.6***)

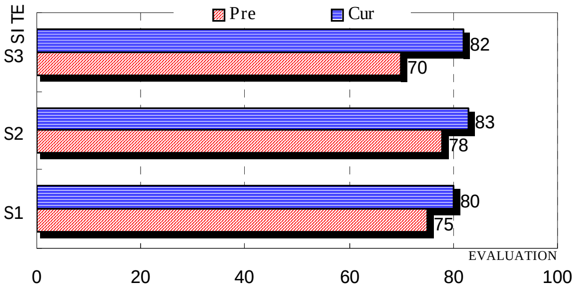

Figure7. Average evaluation scores of 3 learning sites (***V=0.64***)

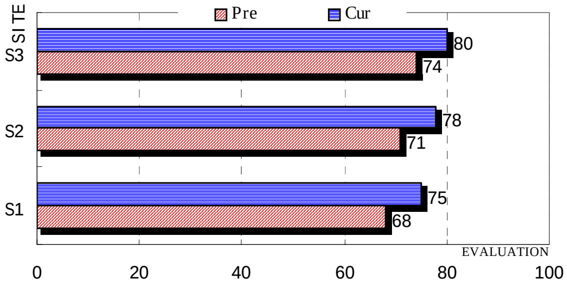

Figure8. Average evaluation scores of 3 learning sites (***V=0.68***)

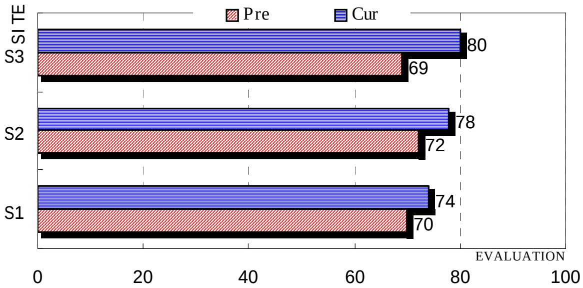

Figure9. Average evaluation scores of 3 learning sites (***V=0.72***)

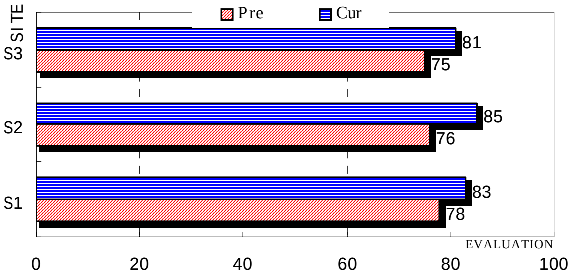

Figure10. Average evaluation scores of 3 learning sites (***V=0.76***)

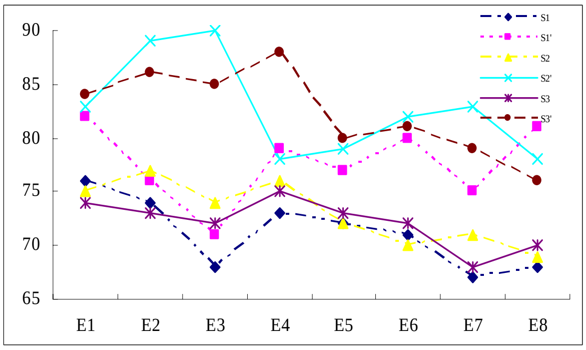

Figure11. Experts reviews on the topic extraction (***V=0.48***)

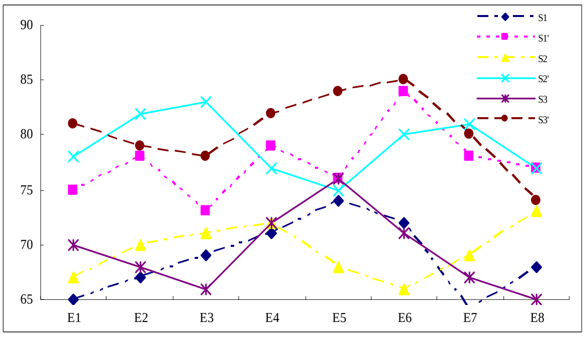

Figure12. Experts reviews on the topic extraction (***V=0.56***)

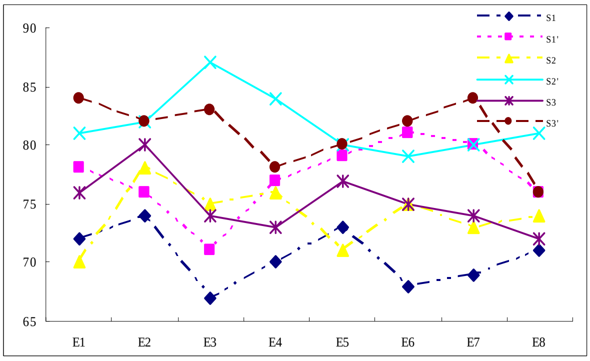

Figure13. Experts reviews on the topic extraction (***V=0.64***)

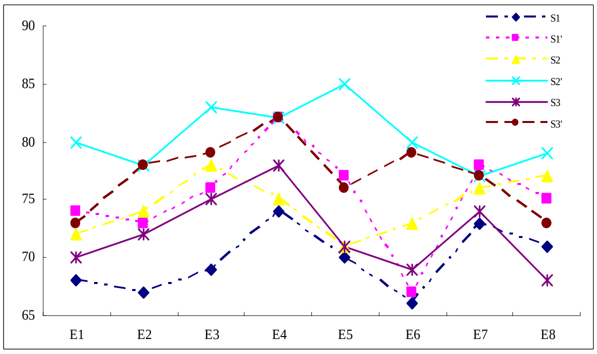

Figure14. Experts reviews on the topic extraction (***V=0.72***)

© 2011 ACADEMY PUBLISHER

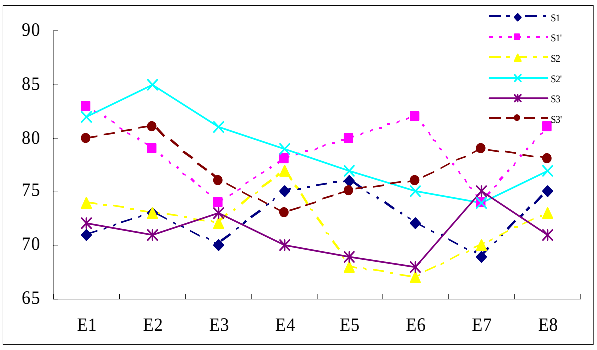

Figure15. Experts reviews on the topic extraction (***V=0.64***)

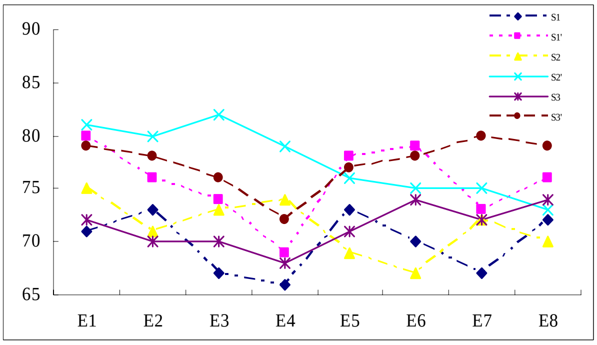

Figure16. Experts reviews on the topic extraction (***V=0.68***)

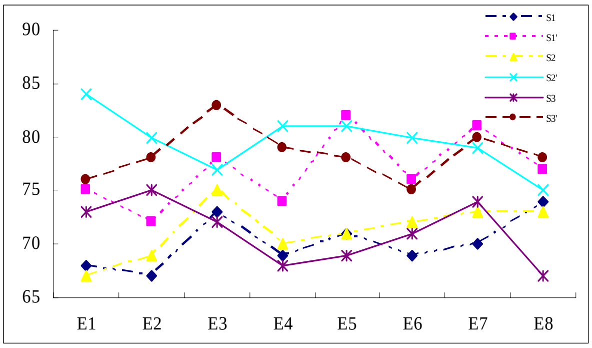

Figure17. Experts reviews on the topic extraction (***V=0.72***)

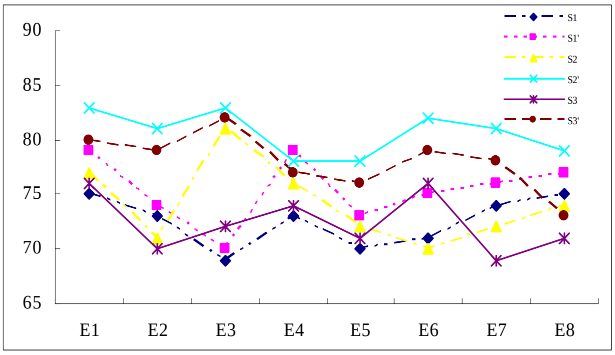

Figure18. Experts reviews on the topic extraction (***V=0.76***)

V. Conclusions and Future Work

The project "Computer System and Interface Technology" funded by the National High Technology Research and Development Program ("863"Program) of China" Research and Development of E-learning Platforms and Education Resource", has carried out the artificial extraction and semantic mark of topic. On this basis, the study of topic extraction method is a feasible new idea. Facing with massive learning resources, we apply the improved OPTIC-based NOP algorithm (DBCO), computing the correlation between topic and the documents in learning resources. The method has been verified to be effective in plain text information retrieval and data mining. We believe that it will also be feasible in multimedia learning resources after being preprocessed. By studying the effective algorithm of topic extraction, and analyzing the limitations of complexity of time and space with a large number of experiments, we verify that the algorithm is effective and it can meet customers' requirements on efficiency and accuracy of topic presentation. This paper's contribution includes: (1) colleting the theme concept rather than the word form, using semantic resources WordNet for word sense disambiguation of words in the learning resources documents, and then extracting the theme concept to build vector space model for topic extraction; (2) proposing a density-based clustering algorithm, applying it into web course systems; It divides multi-document collections into different sentence clusters; then extracts a certain number of sentences from different sub-themes to produce the digest; and (3) sorting the output to generate topic content.

This work could be extended in several directions. Our data sets and experiments are derived from real-world resources. Although the proposed approach requires training data, it may not be necessary to retrain the model frequently. As long as the web learning resources used to compute the class clustering reflect the changes in the sub-themes' relevance to a topic, this method is effective. Future work might include empirically evaluating the robustness of large–scale learning resources derived in a dynamic environment; and binding the user information protection with user's trust links into the topic extraction model.

Acknowledgement

This work is supported by the National High Technology Research and Development Program ("863"Program) of China (Grant No. 2003AA001032); the Fundamental Research Funds for the Central Universities (Grant No. 20102110101000109); the Natural Science Foundation of China (Grants No. 60672051); the Science Research Project of Educational Commission of Guangxi (Grants No. 200911LX541).

References

[1] S.B. Kim , H. C. Seo and H. C. Rim. Information Retrieval using Word Senses: Root Sense Tagging Approach [A]. In : Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval[C]. ACM Press, 2004. 258-265

[2] S. Liu, F. Liu, C. Yu and W. Meng. An Effective

Approach to Document Retrieval via Utilizing WordNet and Recognizing Phrases[A]. In: Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval[C]. ACM Press, 2004. 266-227

[3] S. Liu, C. Yu and W. Meng. WordSense Disambiguation in Queries [A]. In: Proceed2ings of the 14th ACM International Conference on Information and Knowledge Management[C]. ACM Press, 2005. 525-532

[4] M. Ankerst, M. M. Breunig, H. P. Kriegel, J. Sander. OPTICS: ordering points to identify the clustering structure. International Conference on Management of Data [A]. In: Proceedings of the 1999 ACM SIGMOD international conference on Management of data[C]. ACM Press,1999. 49-60

[5] R. Barzilay, N. Elhadad, K. McKeown, Inferring Strategies for Sentence Ordering in Multidocument News Summarization. in Journal of Artifical Intelligence Research (JAIR) [J], 2002, Vol. 17, pp 35-55

[6] C. Y. Lin, C.Y. Lin and E. Hovy. Automated Multi-document Summarization in NeATS. In Proceedings of the Human Language Technology Conference[C]. 2002

[7] Y. Li, Q. Zhong, J. Li, J. Tang, Result of ontology alignment with RiMOM at OAEI'07. In Proc. of International Workshop of Ontology Matching on the 6th International Semantic Web Conference[C].

[8] W. Hu, Y. Zhao, Y. Qu, Partition-based Block Matching of Large Class Hierarchies[C] Proc. of the 1st Asian Semantic Web Conference. Beijing, China, 2006: 72-83.

Ming Xie

Hunan, China
Ph.D. in Computer Science
Computer School of Wuhan University,
Wuhan, Hubei, China
430072

Email: wolfgangtse@gmail.com

Chanle Wu

Hubei, China
Professor of Computer Science
Computer School of Wuhan University,
Wuhan, Hubei, China
430072

Email: wuchle @whu.edu.cn

© 2011 ACADEMY PUBLISHER