SCORM에서 SCO의 클러스터링 기법

윤 홍 원*

A Method of Clustering for SCOs in the SCORM

Hong-won Yun*

요 약

SCORM에서 SCO는 학습자가 검색하는 학습 단위가 된다. e-러닝 환경에서 학습자가 찾는 SCO를 신속하게 검색할 수 있는 저장 방법이 필요하다. 본 논문에서는 SCO의 클러스터링 방법을 수학적으로 정형화하여 정의하였다. 또한 SCO를 평가하는 기준을 제시하였고 각 SCO를 평가하는 절차를 나타내었다. 실험을 통하여 제안한 클러스터링 방법에 기반을 둔 검색이 기존의 검색 방법보다 성능이 우수함을 보였다.

ABSTRACT

A SCO is a learning resource that is retrieved by a learner in the SCORM. A storage policy is required a learner to search SCOs rapidly in e-learning environment. In this paper, We define the mathematical formulation of clustering method for SCOs. Also we present criteria for cluster evaluation and describe procedure to evaluate each SCO. We show the search based on proposed clustering method increase performance than the existing search through performance evaluation.

키워드

SCORM, SCO Evaluation, Cluster-based Search, Probabilistic Clustering

Ⅰ. 서 론

SCORM(Sharable Content Object Reference Model)은 웹 기반 학습 환경의 기술적인 기반을 정의한다. SCORM은 콘텐츠 집합 모델(CAM: Content Aggregation Model)과 실행 환경(RTE: Run-Time Environment)을 정의하고 학습자의 요구에 따라 다양하게 학습 콘텐츠를 제공하기 위한 시퀀싱과 네비게이션을 정의한다[1]. SCORM에서 LMS (Learning Managemetn System)는 학습 콘텐츠를 학습자에게 전달하고 관리하는 역할을 수행한다. LMS가 학습자에게 전달하고 추적할 수 있는 학습의 가장 작은 논리적 단위를 SCO(Sharable Content Objects)라고 한다. SCO는 텍스트, 이미지, 사운드 등과 같은 하나 이상의 Asset으로 구성되며 학습자원을 구성하는 기본 요소이다[2,3,4].

SCORM의 기본적인 학습 순서 모델은 No Sequencing rules, Linear, Linear Choice, Constrained Choice 등이 있다[3]. 현재 SCORM에서 학습 순서에 따라 학습하고 학습 내용을 찾는 방법은 지원하고 있으나 학습자가 학습 내용을 검색하는 경우에 해당하는 SCO를 신속하게 제공하는 방법에 대한 연구 결과는 없다. SCO는 Asset으로 구성되며 Asset은 text이거나 non-text로 구성되는데 Asset의 특성을 활용하고, SCO의 저장 방법을 개선한다면 학습

* 신라대학교 컴퓨터정보공학부

접수일자 : 2006. 9. 27

내용의 검색 속도를 향상시킬 수 있다.

지금까지 SCO의 저장 방법에 대한 연구는 없으나 인접한 연구로써 텍스트 카테고리를 적용한 클러스터 기법에 관한 연구가 있다[5,6,7]. 이 연구는 문서에 나오는 텍스트를 카테고리화하여 클러스터링하고 검색하는데 적합한 저장 방법을 제공하고 있으나 SCO와 같은 e-러닝 콘텐츠의 빠른 검색을 위한 저장 방법으로 적합하지 않다.

본 논문에서는 SCO의 클러스터링 방법을 수학적으로 정형화하여 정의한다. 실험을 통하여 제안한 SCO의 클러스터링 저장 방법이 기존의 방법보다 성능이 우수함을 보인다. 본 논문의 구성은 다음과 같다. 2장에서는 본 연구와 인접한 연구를 소개하고 3장에서는 제안한 클러스터링 방법을 정의한다. 4장에서는 실험 결과를 나타내고 마지막으로 5장에서 결론을 맺는다.

Ⅱ. 관련 연구

2.1. SCO(Sharable Content Object)

SCO는 하나 이상의 Asset으로 구성되며 LMS가 관리하는 학습 단위이다. 각 Asset은 Asset 객체와 함께 메타데이터가 있으며 재사용될 수 있다. SCO는 LMS에 의하여 초기화되고 종료되며 학습 위치를 추적하는 대상이며 학습자가 학습 콘텐츠를 검색하면 검색의 대상 객체가 된다[1,8]. 본 논문에서는 학습 콘텐츠 검색시 대상 객체로써 SCO를 단위로 한다. 최근에는 SCORM의 확장에 관한 연구가 진행되고 있으며 이러한 연구에는 콘텐츠 데이터 모델의 구조, 런타임 환경의 설계 등이 포함되어 있다[1,11].

2.2. 카테고리 검색

텍스트 카테고리화는 카테고리 검색의 한 과정으로 볼 수 있다. 텍스트 카테고리화는 온라인 정보의 급속한 증가로 인하여 텍스트 처리 기술의 핵심이 되었으며 인터넷이나 인트라넷 상의 도큐먼트를 분류하는 데 이용되고 있다. 텍스트 카테고리화의 목표는 주어진 도큐먼트를 이미 정의된 카테고리로 분류하는 것이다. 각 도큐먼트는 기계학습을 통하여 하나 또는 그 이상의 카테고리로 분류된다[5,9,10]. 텍스트 카테고리화와 관련하여 기계 학습에서 도큐먼트의 재표현, 분류 기술, 분류의 평가 등의 연구가 있다. 본 논문에서는 SCO의 클러스터링에 기반을 둔 검색과 카테고리화에 기반을 둔 검색을 비교하기 위하여 카테고리 검색 방법을 비교 연구로 활용한다.

Ⅲ. 클러스터링 기법

본 장에서는 SCO의 클러스터링 기법을 제안한다. 먼저, SCO의 클러스터를 정형화하여 정의하고, SCO의 평가 기준과 평가 절차를 제시한다.

3.1. 클러스터 정의

$S = \{s_1, s_2, ..., s_N\}$ 는 서로 다른 N개 SCO 공간의 집합이라고 한다. SCO 공간의 부분집합 $S_P \subseteq S$로 둔다. 각 Asset은 text 이거나 non-text 이고 text로 표현하는 용어의 집합과 non-text로 된 Asset의 메타 데이터로 사용된 용어의 집합을 이루는 서로 다른 M개 용어의 집합을 T라 하고 $T = \{t_1, t_2, ..., t_M\}$ 으로 한다. T의 부분집합 $S_T \subseteq T$로 둔다.

하나의 SCO는 g로 쓰고 S_T와 S_P 조합으로 정의한다:

$$g = (S_T, S_P) \tag{1}$$

Asset과 SCO 공간의 개수가 각각 M, N 이므로 $M \times N$ 이 행렬 위에 임의의 Asset이 어떤 SCO 공간에 있는지 나타낼 수 있다. $M \times N$ 행렬의 (i, j) 번째 셀은 s_j ($\in$S)에 나타내는 t_i ($\in$T)를 뜻한다.

t_i가 SCO 공간의 집합에 포함된 횟수를 $f(t_i, S)$로 나타낼 수 있고 마찬가지로 s_j에 포함되어 있는 용어의 개수를 $f(T, s_j)$로 나타낼 수 있다. 따라서 $f(t_i, s_j)$는 s_j에 나타나는 t_i의 횟수가 된다. 모든 SCO 공간 S에서 모든 T의 횟수를 $f(T, S)$라고 하면 다음과 같다.

$$f(T, S) = \sum_{t_i \in T} f(t_i, S) = \sum_{s_j \in S} f(T, s_j) = \sum_{t_i \in T} \sum_{s_j \in S} f(t_i, s_j) \tag{2}$$

각 용어 t_i ($\in$T)의 확률은 아래와 같다.

$$P(t_i) = \frac{f(t_i, S)}{f(T, S)} \tag{3}$$

사후 확률(posterior probability) $P(s_j|t_i) = P(t_i, s_j) / P(t_i)$ 이므로,

$$P(t_i, s_j) = \frac{f(t_i, s_j)}{f(t_i, S)} \cdot \frac{f(t_i, S)}{f(T, S)} \quad (4)$$

같은 방법으로 S_T와 S_P에 적용하면 다음과 같다.

$$P(S_T, S_P) = \frac{f(S_T, S_p)}{f(S_T, S)} \cdot \frac{f(S_T, S)}{f(T, S)} \quad (5)$$

non-text로 된 Asset의 메타 데이터로 사용된 용어의 집합을 이루는 서로 다른 L개 용어의 집합을 X라고 하고 $X = \{x_1, x_2, ..., x_L\}$, X의 부분집합 $S_X \subseteq T$로 둔다. 각 non-text로 된 Asset의 x_i ($\in T$)의 확률은 다음과 같다.

$$P(x_i) = \frac{f(x_i, S)}{f(X, S)} \quad (6)$$

non-text로 된 Asset에 같은 방법으로 적용하면 다음과 같다.

$$P(x_i, s_j) = \frac{f(x_i, s_j)}{f(x_i, S)} \cdot \frac{f(x_i, S)}{f(X, S)} \quad (7)$$

S_T와 S_P에 같은 방법으로 적용하면 다음과 같다.

$$P(S_X, S_p) = \frac{f(S_X, S_P)}{f(S_X, S)} \cdot \frac{f(S_X, S)}{f(X, S)} \quad (8)$$

3.2. SCO의 평가

용어와 SCO 공간의 대응하는 이산적인 두 개의 임의 변수를 각각 T와 S라고 하면 확률 이론에서 T와 S 사이에 상호 정보(mutual information)는 $I(T;S)$로 표현하고 다음과 같이 계산한다.

$$I(T;S) = \sum_{t_i \in T} \sum_{s_j \in S} P(t_i, s_j) \log \frac{P(t_i, s_j)}{P(t_i)P(s_j)} \quad (9)$$

여기서 $P(t_i, s_j)$와 $P(t_i)$는 각각 식(4)와 식(3)과 같고, $P(s_j) = f(s_j) / f(T, S)$이다.

non-text로 된 Asset과 SCO 공간의 대응하는 이산적인 두 개의 임의 변수를 각각 X와 S라 하고 X와 S 사이의 상호 정보는 $I(X;S)$로 나타낸다.

$$I(X;S) = \sum_{x_i \in X} \sum_{s_j \in S} \log \frac{P(x_i, s_j)}{P(x_i)P(s_j)} \quad (10)$$

이 식에서 $P(x_i, s_j)$와 $P(x_i)$는 각각 식(7)과 식(6)과 같다. non-text로 된 Asset의 가중치를 ωI로 쓰고 계산하면 $\omega I(T;S) = I(T;S) \times I(X;S)$ 가 된다. 하나의 Asset과 SCO 공간을 조합한 값은 $vI(T;S)$로 쓰고, $vI(T;S) = I(T;S) + \omega I(T;S)$로 계산한다. 따라서 $vI(T;S) = (1 + I(X;S)) \times I(T;S)$ 가 된다.

각 SCO의 상호 정보 값은 다음과 같다.

$$I(S_T;S_P) = P(S_T, S_P) \log \frac{P(S_T, S_P)}{P(S_T)P(S_P)} \quad (11)$$

여기서 $P(S_T) = \sum_{t_i \in S_T} P(t_i)$, $P(S_P) = \sum_{s_j \in S_P} P(S_j)$ 이다.

각 SCO의 값은 $\eta I(S_T;S_P)$로 표현하고 $\eta I(ST;SP) = (1 + I(S_X;S_P) \times I(S_T;S_P))$로 한다. 용어 대 SCO 공간의 일대일 대응에서 k개의 임의 변수 $R_1, R_2, ..., R_k$로 확장하면 다음과 같다.

$$vI(R_1;R_2;...;R_k) = \sum_{i=1}^{k} (I(R_1;R_2;...;R_k) + \omega I(R_1;R_2;...;R_k)) \quad (12)$$

여러 개 Asset으로 구성된 SCO의 평가값 $vI(R_1;R_1;...;R_k)$를 줄여서 svI로 쓰기로 한다.

3.2. SCO 평가의 절차

각 SCO를 평가하는 절차는 다음과 같다. 첫째, 각 Asset을 표현하는 용어를 선택하고 그 용어가 들어갈 SCO 공간을 선택한다. 각 용어는 텍스트로 된 Asset을 대표하는 용어와 non-tex인 Asset을 설명하는 메타 데이터를 포함한다. 각 용어와 용어가 포함되는 SCO 공간은 높이가 2인 그래프로 나타낼 수 있다. 둘째, 각 용어가 대응하는 SCO 공간에 나타나는 빈도를 계산한다. 각 용어의 빈도는 $M \times N$의 2차원 행렬로 나타낸다. 앞에서 보인 식(4)로 각 용어의 확률을 계산한다. 셋째, non-text인 Asset을 가지는 용어의 확률을 식(7)을 적용하여 계산한다. 넷째, 용어의 부분집합이 되는 각 SCO를 평가한다. 각 SCO의 평가는 식(12)를 적용하여 계산한다. 이 식에는 위 셋째 단계에서 구한 non-text인 Asset의 확률을 가중치로 사용한다. non-text가 없는 텍스트인 Asset은 가중치가 0이

된다. 다섯째, 평가된 상위 k개의 SCO를 묶어서 클러스터로 한다.

Ⅳ. 성능 평가

본 장에서는 기존의 카테고리에 기반을 둔 검색과 제안한 클러스터링 저장에 기반을 둔 검색 사이의 성능을 평가한다.

4.1. 실험 환경

본 실험에서는 5,000개의 SCO를 대상으로 하였으며 5,000개의 SCO에서 서로 다른 용어는 1,000개가 있다고 가정하였다. 한 개의 SCO에서 중복되는 용어는 최소 1개에서 최대 5개로 하고 한 개 SCO의 크기는 512KB이고 한 개 용어의 크기는 최대 크기는 10 바이트로 하였다. 하드디스크의 탐색 시간(seek time)은 3.5ms로 하였다. SCO의 수를 변화하면서 처리 시간과 정확도를 측정하였고 주어진 시간에서 처리하는 SCO의 수를 비교하였다.

4.2. 평가 결과

기존의 카테고리에 기반을 둔 검색 방법은 각 그림의 범례에서 Category based search로 나타내고 제안한 클러스터링 저장에 기반을 둔 검색 방법은 CBS(Cluster based Search) with probabilistic clustering으로 표현한다. 본 절에서는 Category baed search를 '기존 방법', CBS with probabilistic clustering을 '제안한 방법'이라고 줄여서 쓰기로 한다.

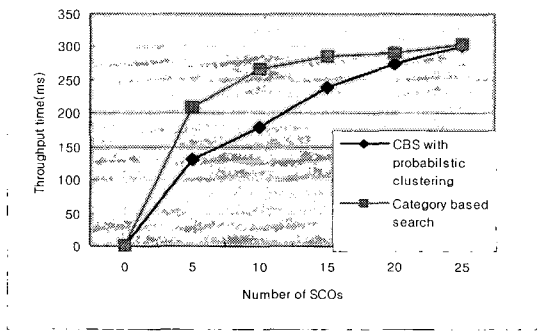


그림 1. 처리 시간 비교
Fig. 1. Comparison of throughput time

그림1은 SCO 수의 변화에 따른 처리 시간의 변화를 나타낸다. SCO의 수가 5개에서 10개 정도에서 제안한 방법의 성능이 우수하였다. 이것은 확률이 높은 용어를 포함한 SCO가 10개 안팎에서 클러스터링되어 있기 때문이다. SCO의 수가 20개 이상이 되면 기존 방법과 제안한 방법의 성능 차이가 비슷하게 되었다. 이것은 20개 정도의 SCO를 검색하면 검색 대상인 용어를 포함한 SCO를 거의 다 읽게 되므로 두 방법은 비슷한 성능을 보이게 된다. 학습자가 검색한 용어를 포함한 SCO가 결과로 제출된 SCO의 수 10개 이하에서 발견된다면 제안한 기법은 성능이 우수하다.

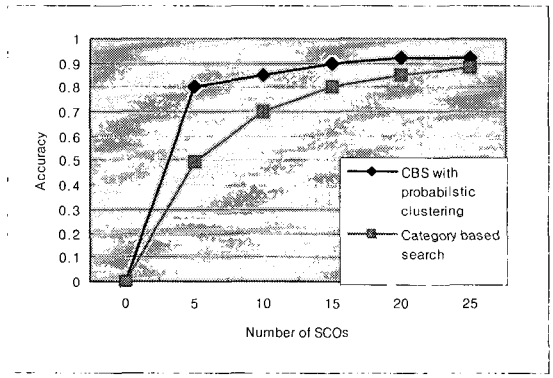


그림 2. 정확도 비교
Fig. 2. Comparison of accuracy

그림2는 기존 방법과 제안한 기법 사이의 정확도를 비교하여 나타낸다. 제안한 방법은 확률이 높고 가중치가 높은 SCO부터 낮은 순으로 결과가 나오고 기존 방법은 카테고리 안에서 임의의 것이 선택되어서 결과로 제출된다. 그림2에서는 검색 결과인 SCO의 수가 적을수록 제안한 방법의 정확도가 높음을 알 수 있다. 제안한 방법은 확률과 가중치가 높은 SCO부터 낮은 순서로 저장되어 있으므로 최초 결과로 제출되는 SCO의 수가 적을수록 정확도가 높고, 기존 방법은 선택된 카테고리 안에서 임의로 저장되어 있기 때문에 선택된 SCO의 수가 적을수록 정확도가 낮다.

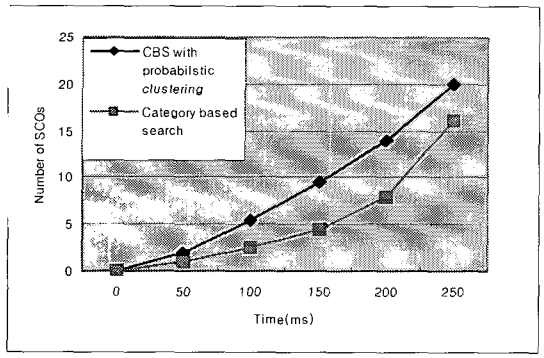


그림 3. SCO의 처리 수
Fig. 3 Number of SCOs

그림 3은 시간의 변화에 따른 SCO의 처리 성능을 나타낸다. 소요 시간이 적을 때 두 기법 사이에 처리한 SCO의 수는 차이를 보이면서 제안한 기법의 성능이 향상되었고 시간이 지날수록 두 기법 사이의 성능 차이는 감소되었다. 검색 초기에는 SCO가 인접하게 저장되어 있는 제안한 기법의 성능이 우수하고 시간이 지나면 기존 방법처럼 다른 SCO를 탐색하는 시간이 필요하므로 두 방법 사이의 격차는 줄어들었다.

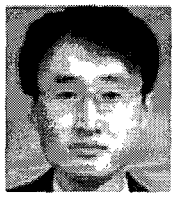
V. 결 론

본 논문에서 학습 콘텐츠의 빠른 검색을 위하여 SCO의 클러스터링 방법을 수학적으로 정형화하여 정의하였다. 또한 SCO를 평가하는 기준과 각 SCO를 평가하는 절차를 제시하였다. 실험을 통하여 기존 방법과 제안한 방법 사이에 처리 시간, 정확도, SCO의 처리 수에 대하여 비교하였다. 제안한 SCO의 클러스터링 저장 방법에 기반을 둔 검색이 기존의 검색 방법보다 성능이 우수함을 보였다. 제안한 클러스터링 방법은 분산 환경의 e-러닝 콘텐츠 저장 방법으로 활용할 수 있다.

참고문헌

[1] SCORM 2004 2nd Edition Overview, p.21, 2004. 07.
[2] SCORM Content Aggregation Model, Version 1.3.1 p.11, 2004. 07.
[3] SCORM Sequencing and Navigation, Version 1.3.1 p.3, 2004. 07.
[4] Gennaro Cstagliola, Filomenna Ferrucci, Vittorio Fuccella, "Web System Architectures: SCORM Run-time Environment as a Service, Proceedings of the 6th International Conference on Web Engineering '06," p.103, 2006. 07.
[5] Makoto Iwayama, Takenobu Tokunaga, "Cluster-based Text Categorization: A Comparison of Category Search Strategies," Proceedings of ACM SIGIR'95, pp.273-278, 1995. 07.
[6] Fabrizio Sebastiani, "Machine Learning in Automated Text Categorization," ACM Computing Surveys, Volume 34 No. 1, pp.1-47, 2003. 03.
[7] Ron Bekkerman, Ran El-Yaniv, Naftali Tishby, Yoad Winter, "On Feature Distributional Clustering for Text Categorization," Proceedings of the 24th ACM SIGIR, pp.146-153, 2001. 09.
[8] SCORM Run-Time Environment, Version 1.3.1 p.3-8, 2004. 07.
[9] Xuanhui Wang, Jian-Tao Sun, Zhen Chen, ChenXiang Zhai, "Machine Learning: Latent Semantic Analysis for Multiple-type Interrelated Data Objects," Proceedings of the 29th ACM SIGIR'06, pp.236-243, 2006. 08.
[10] R. Baeza-Yate, B. Ribeiro-Neto, *Modern Information Retrieval*, Addison-Wesley, 1999.
[11] D. Simoes, R. Luis, N. Horta, "Enhancing the SCORM Metadata Model," Proceedings of the 13th WWW Conference, pp.238-239, 2004. 05.

저자소개



윤 홍 원(Hong-won Yun)

1986년 부산대학교 계산통계학과 졸업(학사)
1990년 한국외국어대학교 경영정보대학원 전자계산학과(이학석사)
1998년 부산대학교 대학원 전자계산학과(이학박사)
2003년 North Carolina State University 객원교수
1996년~현재 신라대학교(구.부산여자대학교) 컴퓨터정보공학부 부교수
※관심분야: 데이터베이스 시스템, 시간 데이터베이스, 시맥틱 웹, e-러닝