다차원 FCM을 이용한 웹 로그 데이터의 유사 패턴 분석

김미라[0] 조동섭
이화여자대학교 과학기술대학원 컴퓨터학과
{memory[0], dscho}@ewha.ac.kr

Similarity Pattern Analysis of Web Log Data using Multidimensional FCM

Mi ra Kim[0] Dong sub Cho
Dept. of Computer Science and Engineering, Ewha Womans University

요 약

데이터 마이닝(Data Mining)이란 저장된 많은 양의 자료로부터 통계적, 수학적 분석방법을 이용하여 다양한 가치 있는 정보를 찾아내는 일련의 과정이다. 데이터 클러스터링은 이러한 데이터 마이닝을 위한 하나의 중요한 기법이다. 본 논문에서는 Fuzzy C-Means 알고리즘을 이용하여 웹 사용자들의 행위가 기록되어 있는 웹 로그 데이터를 데이터 클러스터링 하는 방법에 관하여 연구하고자 한다. Fuzzy C-Means 클러스터링 알고리즘은 각 데이터와 각 클러스터 중심과의 거리를 고려한 유사도 측정에 기초한 목적 함수의 최적화 방식을 사용한다. 웹 로그 데이터의 여러 필드 중에서 사용자 IP, 시간, 웹 페이지 필드를 WLDF(Web Log Data for FCM)으로 가공한 후, 다차원 Fuzzy C-Means 클러스터링을 한다. 그리고 이를 이용하여 샘플 데이터와 임의의 데이터간의 유사 패턴 분석을 하고자 한다.

1. 서 론

인터넷의 발달로 인하여 정보의 중요성이 강조되어지고 있다. 사용자의 다양한 자료를 통해서 의미 있는 정보를 얻기 위해서는 데이터 클러스터링이 많이 이용된다. 클러스터링(clustering)이란 주어진 데이터 집합을 서로 유사성을 가지는 몇 개의 클러스터로 구분해 나가는 과정으로, 하나의 클러스터에 속하는 데이터 점들 간에는 서로 다른 클러스터 내의 점들과는 구분되는 유사성을 갖게 된다. 데이터 마이닝(data mining)에서 클러스터링 방법은 기존의 통계, 기계학습, 패턴인식에서 쓰이던 방법에 부가적으로 데이터베이스 지향적인 제약 사항들(제한된 메모리 양, I/O 시간 최소화 등)을 첨가시킨 것으로서, 최근의 멀티미디어 데이터와 같이 혼합되고 다양한 다차원 데이터를 효율적으로 분류해 나가기 위한 방안으로 연구되고 있다. 사용자가 웹 사이트를 이용하면 이에 대한 기록이 로그라는 형태로 흔적이 남는다. 로그분석이란 이 데이터를 기반으로 다양한 정보를 추출해 내는 것이다. 사용자가 웹 사이트에 접속한 후 모든 작업들은 웹 서버를 통해 이루어진다. 사용자가 웹 서버에 접속을 하게 되면 그 이후의 모든 작업들은 웹 서버에 미리 정해 놓은 위치에 데이터로 남게 된다. 특정 웹 페이지를 보기 위한 사용자의 요구로, 웹 서버는 해당 웹 페이지와 관련된 여러 파일 등에 접근하게 된다. 따라서 사용자가 요청하는 특정 웹 페이지뿐만 아니라 해당 웹 페이지와 관련된 이미지 파일, 이미지 데이터, Include 파일 등에 대한 정보가 로그 파일에 저장된다.

이에 본 논문에서는 Fuzzy C-Means 클러스터링을 이용한 웹 로그 분석 기법을 제안하였다. 웹로그 데이터를 WLDF로 가공한 후, FCM에 적용하여 유사 패턴을 알아보고자 한다. 웹로그 데이터를 요일별 샘플 데이터를 정의한 후, 유사 패턴 분석 방법을 통해 임의의 웹로그 데이터의 요일을 찾아내고자 한다.

본 논문의 구성은 다음과 같다. 제 2장에서는 본 논문이 제안하는 Fuzzy C-Means 클러스터링을 이용한 웹로그 분석의 설계를 보인다. 제 3장에서는 FCM과 WLDF를 이용하여 similar pattern을 분석하였다. 제 4장에서는 결론과 향후연구과제를 기술한다.

2. 다차원 FCM을 이용한 웹로그 분석 기법

본 논문에서는 다차원 FCM을 이용한 웹로그 분석 기법을 제안하고자 한다. 우선 서버에 기록되어 있는 웹로그 데이터에 대한 전처리 과정이 필요하다. 웹로그에서 패턴 발견을 위해 데이터를 추출, 변환, 정제하는 일련의 과정을 전처리 과정이라고 한다. 이러한 전처리 과정이 필요한 이유는 웹로그 데이터를 FCM으로 클러스터링 하기 위해서는 웹로그 데이터가 전처리 과정과 같이 가공되어야 하기 때문이다. 이렇게 전처리 된 웹로그 데이터를 Web Log Data for FCM(WLDF)라고 부르도록 하겠다. 본 논문에서는 요일별 유사패턴을 파악하고자 하기 때문에 서버에 저장되어 있는 웹로그를 일별로 구분하기로 하겠다.

이렇게 가공된 웹로그 데이터 WLDF 데이터들을 우선 요일별로 웹로그 샘플 데이터를 만들어 놓아야 한다. 이들을 각각 MSP, TSP, WSP, ThSP, FSP, StSP, SSP라고 정의한다. 이러한 웹로그 샘플 데이터를 정의하기 위해서는 각각의 요일별 가공된 웹로그 데이터 WLDF를 FCM 알고리즘을 통해 클러스터링한다. 이렇게 클러스터링 된 값은 본 실험에서 샘플 데이터 값이 되는 것이다. 샘플 데이터값을 요일별로 정의 한 후에, 임의의 웹로그 데이터값들을 WLDF로 각각의 웹로그 데이터 값들로 가

이 논문은 2002년도 두뇌한국21사업에 의하여 지원되었음.

공한 후에 FCM 알고리즘을 적용하여 데이터 클러스터링을 한다. 그리고 난 후 데이터 클러스터링 된 결과값들을 샘플 데이터 값들과의 상관관계를 파악한다. 이러한 작업을 통하여 임의의 웹로그 데이터가 어떤 요일의 웹로그 데이터인지를 알 수 있게 된다. 그림 2-1은 전체적인 시스템 흐름도를 나타내고 있다.

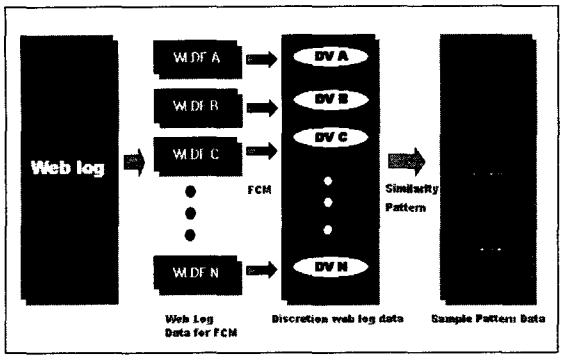

그림 2-1 전체 시스템 구성도

(1) WLDF(Web Log Data for FCM)

WLDF를 만드는 과정은 다음과 같다. Time의 경우는 초 단위를 삭제하고 Float형으로 변환해야 한다. IP의 경우는 사용자를 구분하기 위한 처리로써, 역시 Float형으로 변환해야 한다. 그리고 Web page 정보는 site web page에 임의의 숫자를 부여하여 정수형 데이터로 변환해야 한다. 예를 들면, time의 경우는 04:34:35과 같은 형식의 웹로그 정보를 04.34와 같은 실수형 데이터로 변화하고 초 단위는 삭제를 하여야 한다. Ip의 경우는 203.255.177.177을 교내에서만 접속하는 사용자의 IP를 생각하여 203.255의 단위를 삭제하여 1177.177로 변환하여야 한다. web page 정보의 경우는 웹페이지 정보는 문자 정보이므로, 이를 숫자와 하여서 임의로 저장하고자 한다. 웹로그 데이터의 전처리 과정과 그 예는 다음 그림 2-2와 같다.

time	delete a second. convert into float.
IP	the distinction between user. convert into float.
Web page	assign the option number. convert into integer

time	IP	Web page
04.34	177.177	24
06.67	104.155	15
09.34	198.120	17
10.34	178.34	8
11.23	189.84	5
11.45	157.56	29
11.58	194.138	16
12.45	177.54	12
13.23	104.56	6

그림 2-2 WLDF(Web Log Data for FCM)

(2) FCM(Fuzzy C-means) 알고리즘

FCM 알고리즘은 각 데이타와 각 클러스터 중심과의 거리를 고려한 유사도 측정에 기초한 목적 함수의 최적화 방식을 사용하며, 목적 함수를 다음과 같이 정의된다.

$$Jm(U, V) = \sum_{j=1}^{n} \sum_{i=1}^{c} (u_{ij})^m \, ||x_j - v_i||^2$$

여기서 n은 데이터의 개수, c는 클러스터의 개수이고. 주어진 입력 데이터 집합 X = { $x_1, \cdots, x_n$}에 대한 퍼지 c분할을 $c \times n$의 행렬 U로 나타낼 때 u_{ij}는 데이터 x_j가 클러스터 i에 속하는 소속 정도를 나타낸다. 또한 $\| \cdot \|$은 유클리디안 노름(Euclidean norm)이고, v_i는 i번째 클러스터의 중심을 나타내며, $m \in \{1, \cdots \infty\}$은 퍼지 정도를 나타내는 매개변수이다.

FCM 알고리즘의 수행절차는 다음과 같다.

단계 1: $c(2 \le c \le n)$ 값과 $m(1 \le m \le \infty)$값을 결정한다.

단계 2: 다음의 조건을 만족하는 퍼지 c 분할 $U^{(0)}$을 초기화한다.

$$\sum_{i=1}^{c} u_{ij} = 1, 0 < \sum_{j=1}^{n} u_{ij} < n; u_{ij} \in [0,1], 1 \le i \le c, 1 \le j \le n$$

단계 3: 각 클러스터에 대한 클러스터의 중심 $v_i^{(l)}$를 구한다. (l=0,1,2, · · ·).

$$v_i^{(l)} = \frac{\sum_{j=1}^{n} (u_{ij}^{(l)})^m x_j}{\sum_{j=1}^{n} (u_{ij}^{(l)})^m}, \quad 1 \le i \le c$$

단계 4: 구해진 $v_i^{(l)}$를 이용하여 $U^{(l+1)}$을 계산한다.

$x_j \neq v_i^{(l)}$인 모든 $i \in N_c$에 대하여,

$$u_{ij}^{(l+1)} = \frac{1}{\sum_{k=1}^{c} \left(\frac{||x_j - v_i^{(l)}||}{||x_j - v_k^{(l)}||} \right)^{2/(m-1)}}, 1 \le i \le c, 1 \le j \le n$$

$x_j = v_i^{(l)}$인 어떤 $i \in I \subset N_c$에 대하여,

$\sum_{i \in I} U_{ij} = 1, i \in N_c - I$ 인 경우에 대하여 $u_{ij} = 0$,

단계 5: 만약 $||U^{(l+1)} - U^{(l)}|| \le \varepsilon$ 이면 알고리즘을 끝내고, 그렇지 않으면 단계 3으로 간다.

3. FCM을 이용한 웹로그 분석 기법의 구현 및 평가

실험 결과값의 각 필드에 대한 설명은 그림 3-1과 같다. TI(Total Iteration)는 FCM 알고리즘을 실행하여 최종 데이터 클러스터링까지 걸린 총 반복횟수를 의미한다. V는 FCM 알고리즘 실행 후 최종적으로 얻어지는 중심값이다. 이 세 개의 중심값을 중심으로 데이터 클러스터링이 된다. DV는 FCM 알고리즘을 수행하는데 있어 초기 센터값에서 최종센터값을 뺀 값이다. MD(Moving Distance)는 similarity pattern 분석을 위해 필요한 값으로, 초기센터값에서 최종센터값까지 변한 거리값이다. mean_e와 max_e는 FCM 알고리즘의 변수값으로 이 두 값이 각 알고리즘 수행에 있어서 최소가 될 때까지 반복을 계속한다.

TI	The total iteration number	
V	The final clustering center	(V1,V2,V3)
DV	The initial center - the final center (SV - V)	(DV1,DV2,DV3)
MD	The moving distance from the initial center to the final center	(MD1,MD2,MD3)
mean_e	The iteration should keep up until it was at a minimum	the variable in FCM
max_e	The iteration should keep up until it was at a minimum	the variable in FCM

그림 3-1 Experiment values

월요일부터 일요일까지의 요일별 웹로그를 이용하여 우선 샘플 데이터값을 구해야 한다. 월요일부터 일요일까지의 각각의 웹로그 데이터에 대한 WLDF를 만든후, FCM 알고리즘을 실행한 결과는 다음 그림 3-2과 같다. 각각의 요일별 샘플 데이터 값은 MSP, TSP, WSP, ThSP, FSP, StSP, SSP라고 정의한다. MSP의 (MD1, MD2, MD3) 값은 (53,41,77)이다. TSP는 (35,49,13)이고, WSP는 (50,45,11), ThSP(44,40,8),FSP는 (27,34,40), StSP는 (60,78,5), SSP는 (91,58,73)이다. 임의의 웹로그 데이터를 WLDF 전처리 과정을 거쳐 FCM 클러스터링 알고리즘을 실행한 (MD1, MD2, MD3)값을 구한다. 그리고 그 결과값과 샘플 패턴 데이터의 (MD1, MD2, MD3)값과의 상관관계를 통해 임의의 웹로그 데이터의 요일을 알아낼 수 있다.

4	63,65,62	100,98,68	100,96,33	0.0522	0.2070	-13,-35,37	-10,-38,11	-50,-36,46			MSP
24	39,44,29	104,95,48	54,64,67	0.0002	0.0010	10,-14,30	-14,-35,31	-4,-4,12			TSP
17	54,71,30	106,93,54	42,66,80	0.0002	0.0009	-4,-41,29	-16,-33,26	7,-5,6			WSP
11	53,66,34	122,44,61	27,46,54	0.0002	0.0009	-3,-36,25	-32,15,18	6,-6,0			ThSP
13	47,74,21	36,44,38	27,46,54	0.0003	0.0007	-7,15,27	-42,37,9	7,-6,0			FSP
30	51,34,40	70,27,64	35,92,78	0.0042	0.0080	-15,-40,7	-39,50,70	9,-10,-25			StSP
3	100,78,25	98,101,40	93,101,36	0.0856	0.2477	-50,-68,34	-8,-41,39	-43,-41,43			SSP

그림 3-2 Sample pattern data

4	50,60,24	101,98,68	99,96,31	0,0583	0,2074	0,-30,35	-11,-38,11	-49,-36,46			MSP
8	34,54,35	114,90,39	94,99,34	0,0,131	0,0613	15,-24,25	-24,30,40	-44,-39,45			SSP
9	48,62,29	106,91,50	47,70,80	0.0002	0.0010	1,-32,30	-16,-31,29	2,-10,0			WSP
17	42,57,32	99,97,68	109,94,31	0.0002	0.0008	7,-27,27	-9,-37,11	-59,-34,48			MSP
13	54,65,30	106,91,49	45,69,78	0.0001	0.0008	-4,-35,29	-16,-31,32	4,-9,1			WSP
16	40,55,33	99,97,68	110,95,18	0.0001	0.0007	9,-25,26	-9,-37,12	-60,-35,61			MSP
11	47,68,35	120,46,64	44,66,81	0.0002	0.0008	2,-38,24	-30,13,15	5,-6,-1			ThSP
18	40,50,44	99,96,68	110,93,19	0.0001	0.0008	9,-25,26	-9,-36,13	-60,-33,60			MSP
17	34,54,35	102,98,25	92,101,41	0.0002	0.0008	15,-24,24	-12,-30,54	-42,-41,38			SSP
9	37,50,36	104,65,43	61,68,65	0.0032	0.0167	12,-20,23	-14,-35,33	-11,-8,14			TSP
9	49,58,29	107,87,47	48,70,80	0.0001	0.0006	0,-28,30	-17,-27,32	1,-10,0			WSP

그림 3-3 Similarity pattern data result

실험한 웹로그 데이터의 수는 총 11개이며, 역시 FCM 알고리즘의 초기값은 V1 = (50.0 , 30.0 , 60.0), V2 = (90.0 , 60.0 , 80.0), V3 = (50.0 , 60.0 , 80.0)으로 실험하였다. 실행한 결과는 아래 그림 3-3과 같다. Data A의 (MD1, MD2, MD3) 값은 (47,41,78)으로 MSP 그룹에 속하며 월요일 웹로그 데이터 였음을 알 수 있다. 이런 식으로 하여 MSP에 속하는 웹로그 데이터는 Data A (47,41,78), Data D (39,40,91), Data F (37,40,93), Data H (27,39,91)이다. TSP에 속하는 웹로그 데이터는 Data J (34,50,20)이며, WSP에 속하는 웹로그 데이터는 Data C (44,46,10), Data K (41,46,10), Data E (46,47,10)이다. ThSP에 속하는 웹로그 데이터는 Data G (45,36,9)이다. SSP에 속하는 웹로그 데이터는 Data I (38,68,73), Data B (38,50,75)이다.

4. 결론 및 향후 연구과제

본 논문에서는 각 데이터와 각 클러스터 중심과의 거리를 고려한 유사도 측정에 기초한 목적 함수의 최적화 방식을 사용하는 Fuzzy C-Means 클러스터링 알고리즘을 사용하여, 웹 로그 분석 기법을 제안하였다. 웹 로그 데이터는 사용자들의 행위정보를 담고 있어서 데이터 마이닝에 유용하다. 이에 웹 로그 데이터를 WLDF와 FCM를 이용하여 다차원 분석을 하였다. WLDF 전처리 과정을 통하여 웹 로그 데이터의 여러 필드 중에서 사용자의 IP, 시간, 웹 페이지 정보를 FCM에 적용하였다. 이러한 과정을 거쳐, 요일별 샘플 패턴을 정의한 후, 유사 패턴 파악을 통하여 임의의 웹 로그 데이터들의 요일을 도출하는 것을 보였다.

본 연구를 통하여 웹 사용자들의 행위정보를 데이터 클러스터링 알고리즘을 이용하여 다차원 분석을 할 수 있었다. 웹 로그 데이터를 FCM을 통하여 새로운 웹 로그 분석 기법을 제안하였다. 본 논문의 결과를 기반으로 데이터 마이닝의 하나의 모듈로 자동화하는 방법, 클러스터링 되어지는 데이터 값들을 비쥬얼하게 보이도록 하는 방법 등에 대한 연구가 필요한 것으로 보인다.

참고문헌

[1] 김미라, 곽미라, 조동섭, "Fuzzy C-Means 클러스터링을 이용한 웹 로그 분석기법,"정보과학회 춘계 학술대회:인간과 컴퓨터 상호작용 제 29권 제 1호, 2002년.

[2] 박승수 이상호, 용환승, 김현희, 최지영, "데이타마이닝 알고리즘의 분류 및 분석," 정보과학회논문지 : 데이타베이스 제28권 제3호, 2001년.

[3] Mi-Ra Kim, Dong-sub Cho, "System Log Analysis Technique using Fuzzy C-Means Clustering," The International Conference on Electrical Engineering, Jeju Island, South Korea, July 2002.

[4] Tian Zhang, Raghu Ramakrishnan, and Miron Livny, "BIRCH : An Efficient Data Clustering Method for Very Large Databases," In Proc. of the ACM SIGMOND Conference on Management of Data, Montreal. Canada, pp.103-114, June 1996.