한국정보전자통신기술학회논문지(jkiiect)'18-12, Vol.11 No.6
논문 18-11-06-658 http://dx.doi.org/10.17661/jkiiect.2018.11.6.658

콘텐츠 큐레이션 플랫폼 성능평가 알고리즘

최 종 호*

Performance Evaluation Algorithm of Contents Curation Platform

Jong-Ho Choi*

요 약 본 논문에서는 콘텐츠 큐레이션 플랫폼의 성능평가 알고리즘을 제안하였다. 성능평가는 초기조건과 종료조건을 설정하고, 4개의 파라미터를 측정하여 수행하였다. 성능평가에 사용되는 파라미터는 서비스 응답속도, 응답결과 정확도, 분석결과 정밀도, 분석결과 재현율이다. 제안한 성능평가 알고리즘의 유효성을 확인하기 위해 118,738건의 데이터베이스를 직접 구축한 후, 인터넷에서 콘텐츠를 오픈하여 이를 클릭한 사용자 행위를 기반으로 실험을 수행하였다. 실험에 사용된 페이지 뷰는 6,975,365건이었다, 실험결과, 4개의 파라미터를 이용하면 콘텐츠 큐레이션 플랫폼의 종합평가와 단위평가가 객관적으로 수행될 수 있음을 확인하였다.

Abstract In this paper, a performance evaluation algorithm of contents curation platform is proposed. The evaluation starts with setting the initial-final conditions, and measures four performance evaluation parameters. These parameters are response rate of service, accuracy of response result, precision of analysis result, and recall rate of analysis result. To verify the effectiveness of the proposed algorithm, a database of 118,738 cases was constructed and the contents was opened on the Internet. In this experiment, 6,975,365 pageviews were applied to analysis user behavior. Through the measurement of four parameters, it was confirmed that the integrated and unit evaluation of content curation platform can be performed objectively.

Key Words : contents, curation, parameter, performance evaluation, platform

1. 서 론

콘텐츠 사용자의 입장에서는 특정 사이트가 양질의 콘텐츠를 제공하는 것에 매력을 느끼지만, 사용자 자신이 선호하는 콘텐츠를 접했을 때 그 효용가치를 강렬하게 느끼게 된다. 이러한 측면에서 소비자가 선호하는 콘텐츠는 물론 그동안 자신이 모르고 있었던 취향을 찾아 새로운 콘텐츠를 추천하는 콘텐츠 큐레이션 서비스가 다양한 형태로 인터넷상에서 제공되고 있다.

콘텐츠 큐레이션에서의 핵심은 소비자 개인의 특성과 관련된 개인정보를 파악하는 것이다. 그러나 개인정보의 직접적인 획득은 회원가입 단계에서 일부 가능하나, 개인정보보호법 때문에 제한적일 수밖에 없다. 따라서 현재의 콘텐츠 큐레이션 시스템은 콘텐츠 검색 및 감상 과정에서 일어나는 사용자의 행동특성으로부터 데이터를 수집하고 분석하는 방향으로 개발되고 있다[1,2,3].

콘텐츠 큐레이션의 중요성이 부각됨에 따라 대규모 콘텐츠를 체계적으로 분류할 수 있는 환경을 구축하고, 사용자가 집중하고 있는 콘텐츠의 형태 및 성격, 내용, 사용자 환경 등 환경특성 및 활동정보를 실시간으로 수집 분석하여 수요자 맞춤형 콘텐츠를 제공하기 위한 콘텐츠 큐레이션 플랫폼을 개발하는 연구가 다수 진행되고 있다[4]. 일부 플랫폼은 상용화의 형태로 제공되고 있으나, 큐레이션 성능평가에 관한 구체적인 연구가 진행되지 않아 콘텐츠 제공자 자체의 주관적인 품질측정 척도를 이용해서 성능을 평가하고 있다.

*Department of IoT Electronic Engineering, Kangnam University(jhchoi@kangnam.ac.kr)
Received September 18, 2018 Revised October 07, 2018 Accepted October 27, 2018

따라서 본 논문에서는 서비스 응답속도, 응답결과 정확도, 분석결과 정확도, 분석결과 재현율을 기반으로 콘텐츠 큐레이션 플랫폼의 성능을 정량적으로 평가하는 알고리즘을 제안하였다. 제안 알고리즘에는 성능평가 초기 및 종료 조건과 가중치 분배 방식을 새롭게 적용하되었다.

제안된 알고리즘의 유용성을 확인하는 것을 목표로 10만건 이상의 콘텐츠에서 최소 300만건 이상의 사용자 활동로그를 수집하는 것을 조건으로 성능평가 파라미터 값을 산출하는 실험을 수행하였다.

실험에서는 118,738건의 콘텐츠에서 6,975,365건의 활동로그를 사용하였다. 실험 결과, 평가된 4개의 파라미터를 활용하면 콘텐츠 큐레이션 플랫폼의 성능평가를 객관적으로 수행할 수 있음을 확인하였다. 향후의 연구방향은 본 논문을 기반으로 콘텐츠 큐레이션 플랫폼의 성능평가 표준화 문서를 개발하는 것이다.

2. 콘텐츠 큐레이션 플랫폼

콘텐츠 큐레이션 플랫폼은 일반적으로 4개의 엔진으로 구성된다. 대규모의 콘텐츠를 분류/저장/관리하는 CME(Contents Management Engine), 실시간으로 CME를 통해 관리되는 콘텐츠의 내용 및 유형, 사용자 환경에서 수집되는 사용자 행위 및 형태, 사용자의 콘텐츠 집중도 등을 분석하는 CDSA(Collection Data Streamming Analysis Engine), 사용자의 활동로그, 개인화 정보, 플랫폼 정보 등 사용자 특성과 관련되는 정보를 수집하는 UIC(User Information Collection Engine), 사용자 단말기 플랫폼과 이기종 시스템 관련 API를 제공하는 MCDE(Multi-Platform Contents Display Engine) 등이다. 콘텐츠 큐레이션 플랫폼의 구성도를 그림 1에 나타냈다.

콘텐츠 큐레이션에서의 기본은 콘텐츠를 관리하는 것이다. CME를 기반으로 하는 CMS(Contents Management System)는 대규모의 콘텐츠를 관리하는

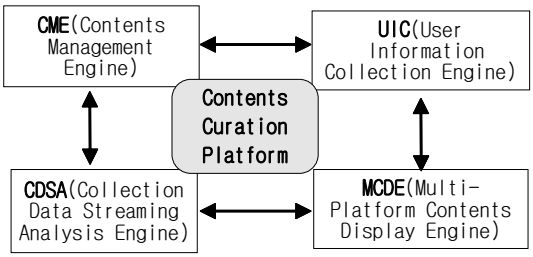

그림 1. 콘텐츠 큐레이션 플랫폼 구성도

Fig. 1. Contents curation platform configuration

시스템으로 콘텐츠 관리와 더불어 콘텐츠별 구매 정산관리, 이벤트 및 설문조사 항목 관리, 사용자 및 발행인 관리, 댓글관리, 맞춤형 시스템 관리 등 다양한 관리 기능을 수행하기 위한 시스템이다. 일반적으로 Saas(Software as a service) 기반으로 구축된다.

콘텐츠 큐레이션에서의 첫번째 단계는 사용자의 환경특성과 활동정보를 실시간으로 수집하는 부분이다. 일반적으로 정보수집 시스템은 Apache의 오픈소스 분산 쿼리 및 처리 엔진인 Spark를 사용하여 구축되고 있다. 시스템 구성도의 예를 그림 2에 나타냈다. 구문 분석기(Phrase Extractor)는 정보수집과 동시에 정보를 분석하는 것을 목표로 적용되는 부분이다. 데이터베이스 인터렉션에서 SQL을 사용하지 않는 데이터베이스 NoSQL 프로젝트 중 하나가 MongoDB이다. MongoDB는 데이터를 저장하는 오프소스 문서지향 데이터베이스이다.

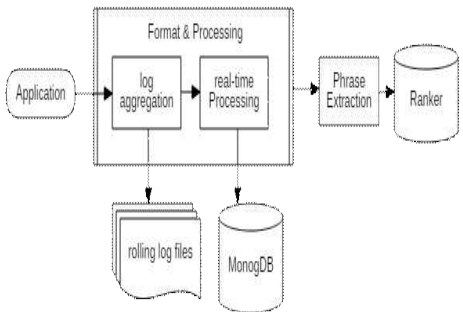

그림 2. 사용자 정보 수집시스템

Fig. 2. User information collection system

콘텐츠 큐레이션 알고리즘은 일반적으로 행동분석 및 내용분석 엔진에 의해서 수행된다. 사용자가 시스

템에 접근하면 사용자의 로그 데이터로부터 얻을 수 있는 환경특성 정보와 사용자가 집중하고 있는 콘텐츠의 형태 및 성격, 내용 등 사용자의 활동 데이터로부터 얻을 수 있는 행동특성 정보를 인덱서가 분류하여 행동분석 엔진 또는 내용분석 엔진으로 전송한다. 행동특성 분석 엔진으로 전송되는 정보는 사용자 쿠키내 키워드와 조회 콘텐츠 등이고, 내용분석 엔진으로 전송되는 정보는 실시간 검색어와 실시간 조회 콘텐츠 등이다. 이 과정에서 내용분석 엔진에서는 콘텐츠 저작자가 등록한 콘텐츠 특성을 참조한다. 행동분석 엔진과 내용분석 엔진에서 추천한 결과의 유사도를 분석함으로써 최종적으로 콘텐츠를 추천하고, 그 결과를 사용자에게 피드백한다. 콘텐츠 큐레이션 알고리즘을 그림 3에 나타냈다.

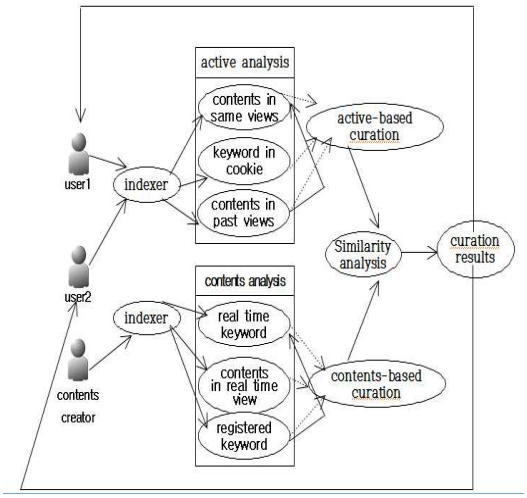

그림 3. 콘텐츠 큐레이션 알고리즘
Fig. 3. Contents curation algorithm

3. 플랫폼 성능평가 알고리즘

현재 콘텐츠 큐레이션 플랫폼이 다양한 분야에서 상업용으로 개발되어 사용되고 있다. 그러나 표준화된 플랫폼의 성능평가 기준이 없어 플랫폼 사용 소비자는 물론 관련 산업계에서도 주관적으로 소프트웨어의 품질을 평가하고 있는 실정이다.

본 논문에서는 플랫폼의 품질을 객관적이고 정량적으로 평가하는 성능평가 알고리즘을 제안하였다. 그림 4에 콘텐츠 큐레이션 시스템의 성능평가 알고리즘을 나타냈다.

시스템 성능평가의 초기조건과 종료조건은 초기 환경설정 및 성능평가 종료에 관한 기본적인 부분이다. 초기조건의 주요 내용은 최소 10만건 이상의 콘텐츠를 대상으로 활동로그가 최소 300만건이상 수집되어야 하고, 별도의 분산처리 없이 서비스가 운영되어야 한다는 것이다. 성능평가 종료조건은 JMeter를 통한 서버 응답속도 측정은 10회 평균값을 200TPS 이상의 기준으로 측정하고, 각 파라미터 측정값이 설정된 기준값을 초과할 경우 종료하는 조건이다.

제안된 성능평가 기준에서 사용되는 파라미터는 서비스 응답속도, 응답결과 정확도, 분석결과 정밀도, 분석결과 재현율이다. 4개의 파라미터를 표 1에 나타냈다[5,6,7,8].

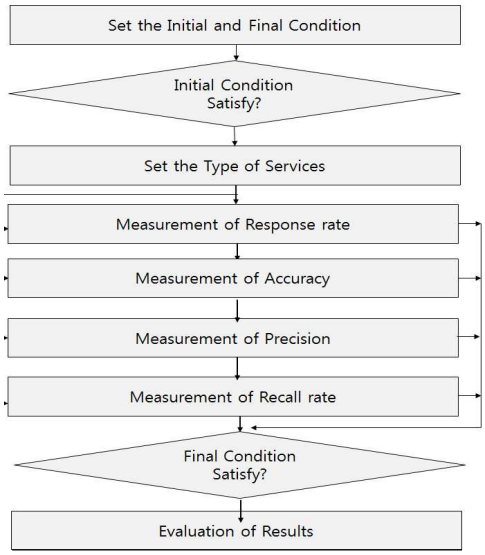

그림 4. 성능평가 알고리즘
Fig. 4. Performance evaluation algorithm

서비스 응답속도는 Apache Jmeter TPS 측정법을 사용하여 200 TPS(Transaction Per Second) 이상에서 측정한다. 그림 5에 Apache Jmeter 측정법을 나타냈다.

표 1. 콘텐츠 큐레이션 플랫폼의 성능평가 파라미터
Table 1. Performance evaluation parameters in contents curation platform

Parameter	Unit	condition
Response rate of services	ms_200TPS	10 times Average
Accuracy of response results	%_accuracy	More than 100,000 cases
Precision of analysis results	precision	More than 100,000 cases
Recall rate of analysis results	recall	More than 100,000 cases

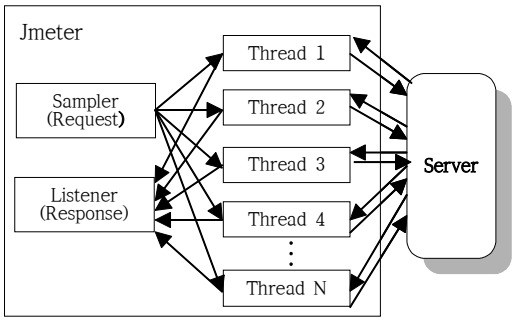

그림 5. Apache Jmeter TPS 측정법
Fig. 5. TPS measurement in Apache Jmeter

응답결과 정확도는 질의어에 대한 응답 정확도로 10만건 이상의 데이터에서 기준 질의어 30개를 선정하는 방식으로 측정한다. 응답결과 정확도는 데이터 셋을 구성하고 질의어 기준으로 정합율을 산정한다.

분석결과의 정밀도 및 재현율은 10만건 이상을 기준으로 각각 식 (1)과 (2)에 의해 산정한다.

$$preision = \frac{N_{hr}}{N_{hr} + N_{mr}} \quad (1)$$

여기서 N_{hr}는 hit 추천수이고, N_{mr}는 miss 추천수이다.

$$recall = \frac{N_{tr}}{N_{tr} + N_{nr}} \quad (2)$$

여기서 N_{tr}는 추천수이고, N_{nr}는 미추천수이다.

본 논문에서는 콘텐츠 큐레이션 플랫폼을 평가하기 위한 4개의 파라미터인 서비스 응답속도, 응답결과 정확도, 분석결과 정밀도, 분석결과 재현율의 가중치를 설정함으로써 플랫폼의 종합평가를 분야별로 실시하는 알고리즘을 제안하였다. 가중치를 성능평가에 적용하는 방법은 서비스 유형(TOS : Type Of Services)을 정의하는 것이다. 표 2에 성능평가 파라미터 기반 서비스 유형을 나타냈다.

표 2. 성능평가 서비스 유형
Table 2. Type of services in performance evaluation

Item No.	TOS Bits	Description
0	0000	Normal(default)
1	0001	Minimize response rate
2	0010	Maximize accuracy
3	0100	Maximize precision
4	1000	Maximize recall

콘텐츠 큐레이션 시스템에서의 성능평가는 종합평가 및 단위별 평가로 구분하여 수행할 수 있다. 종합평가는 4개 항목에서 기준치를 초과할 경우를 대상으로 각 파라미터 값들의 비교를 통해 평가가 가능하고, 단위별 평가에서는 기준치에 관계없이 TOS 1~4 기준에 의거 각각의 항목에 대한 집중적인 성능평가를 진행할 수 있다.

4. 실 험

본 논문에서는 성능평가 알고리즘의 유효성을 확인하기 위한 실험을 수행하기 위하여 상용화 시스템으로 개발한 플랫폼을 사용하였다. 그림 6에 상용 콘텐츠 큐레이션 플랫폼을 나타냈다. 클라우드 기반의 시스템으로 User Information Collector, Data Streaming Filter, CMS, Contents Display System, API 등으로 구성되는 시스템이다.

실험에서 사용하는 플랫폼은 10만건 이상 규모의 빅데이터에서 데이터를 추천하는 시스템이기 때문에 매체 대행사로부터 구매한 118,738건의 빅데이터를 데이터베이스로 구축하였다. 데이터베이스로 구축한 콘텐츠의 일부를 그림 7에 나타냈다.

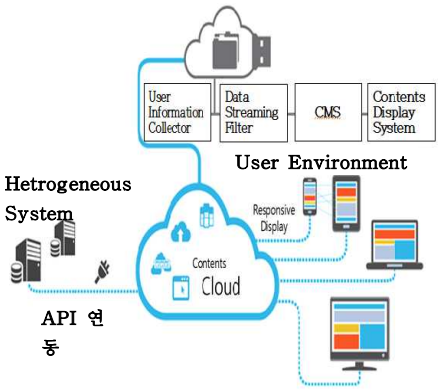

그림 6. 상용 콘텐츠 큐레이션 플랫폼
Fig. 6. Commercial contents curation platform

2017년 1월부터 5월까지 일반 사용자를 대상으로 인터넷상에서 콘텐츠를 오픈하고, 이를 클릭한 사용자 행위를 기반으로 실험을 수행하였다. 총 페이지 뷰 기준 6,975,365건이 발생하였다. 특정 문화사의 디지털 콘텐츠 팀을 통해 “설리의 19금 인스타그램”, “혼전 계약서 쓰는 여자”, “안면 윤곽 주사의 진실” 등 73개의 콘텐츠 각각에 해당하는 추천 콘텐츠를 총 30개 단위로 구성하였다. 73개의 실험 대상 콘텐츠와 “셜리의 19금 인스타그램”에 해당하는 추천 콘텐츠의 일부를 각각 〈표 3〉과 〈표 4〉에 나타냈다.

ar_keyword	ar_title
스트릿, 런던, 패션, 스타일링	TASTE YOU
일본여행, 여행, 일본, 에디터, 사가와, 나고야, 오사카	막내가 간다
팬톤, 컬러, 인테리어, 공간	2017 Spring Color Trend
책,불프강틸만스, 디터람스, 무인양품, 슈퍼노말, 에드루샤, 1984, 무라카미…	벽장을 위한 책
오늘의패션, 패션, 패션화보, 가방, 핸드백, 토트백, 디올, 2017SS, 스타일…	TRUE LADY
오늘의인물, 인물, 인터뷰, 가수, 뮤지션, 김필, 로이킴, 페이스북라이브, 페…	FROM FEEL TO ROY
오늘의신상, 뷰티, 메이크업, 블러셔, 과즙상	봄엔 블러셔
	WHITE AVENUE
남주혁, 청춘, 배우, 드라마, 역도요정김복주, 정준형, 이성경, 로맨스, 신인…	청춘 한가운데, 남주혁

그림 7. 데이터베이스 콘텐츠
Fig. 7 Database contents

콘텐츠 큐레이션 플랫폼의 성능평가 파라미터 중에서 서비스 응답속도를 총 10회 측정한 후, 평균값을 구했다. 서비스 응답속도 측정값을 〈표 5〉에 나타냈다.

표 3. 실험대상 콘텐츠
Table 5. Contents applied in experiments

No.	Contents
1	설리의 19금 인스타그램
2	혼전 계약서 쓰는 여자
∫	∫
72	Poolside Glow
73	차세대 패션 아이콘

표 4. "설리의 19금 인스타그램" 연관 추천 콘텐츠
Table 4. Curation contents associated "Sully Instagram"

No.	기사번호	기사제목
1	30309	돌아온 여자친구
2	29959	송종국, 박잎선 부부, 쌍방 불륜설의 진실
∫	∫	∫
29	31313	방송인 A양의 스폰서 꼬시기
30	30143	정준하 방배동 25억 빌라 매입

표 5. 서비스 응답속도
Table 5. Service response rate

Turn	Minimum arrival time (msec)	Maximum arrival time (msec)	Average time (msec)
1	107	259	183
2	101	231	166
∫	∫	∫	∫
9	105	227	166
10	102	247	175
평균	104.20	254.30	179.20

동일한 조건에서 응답결과 정확도, 분석결과 정밀도, 분석결과 재현율을 측정한 결과는 각각 0.90, 0.62, 0.38으로 측정되었다. 측정결과, 실험 대상 플랫폼의 경우 현재의 상용시스템에서 제시하고 있는 기준값을 초과하는 것으로 나타났다. 본 논문에서 제안한 성능평가 파라미터는 정량적인 측정이 가능하므로 플랫폼을 객관적으로 평가할 수 있다는 특징이 있다. 딥러닝 등의 적용을 통해 큐레이션 시스템의 성능이 지속적으로 높아지고 있으므로 플랫폼의 질적 수준 평가는 상대적인 개념으로 수행되고 있다.

실험을 통해 본 논문에서 제안한 4개의 파라미터를 서비스 유형에 따라 측정한 값을 활용하면 플랫

폼의 성능평가가 다양하게 수행될 수 있다는 것을 확인하였다. 현재 콘텐츠 큐레이션 플랫폼의 성능 인증은 고가의 소프트웨어 인증에 의해서 개발자가 제시하는 조건과 파라미터를 기반으로 수행되고 있다. 본 논문은 성능평가 파라미터 도출, 측정법 제안, 측정 알고리즘의 체계화를 통해 공인인증 표준문서를 개발하는 연구이다.

5. 결 론

콘텐츠 추천에서의 가장 중요한 이슈는 소비자의 행동특성을 수집하는 것이다. 최근들어 콘텐츠 소비자의 행동특성을 수집하고 분석하는 콘텐츠 큐레이션 시장이 급격하게 성장함에 따라 다양한 플랫폼이 상업용으로 출시되고 있다. 그러나 플랫폼의 성능평가에 관한 구체적인 표준이 없기 때문에 플랫폼 시장에서 혼란이 야기되고 있다.

따라서 본 논문에서는 플랫폼의 성능평가 알고리즘을 제안하였다. 상용으로 개발한 플랫폼을 기반으로 성능평가를 수행하고, 성능평가 각각의 파라미터에 대한 객관성을 확인하였다. 콘텐츠 큐레이션 플랫폼 인증체계 구축을 위한 표준문서 작성이 다음 연구과제이다.

REFERENCES

[1] Yeo-Kwang, Yoon, "A Study on Contents Curation of Portal Sites", Journal of the Korea Entertainment Industry Association, Vol. 8, No. 4, 2014.

[2] Sumi Song and Yongik Yoon, "Contents Curation model for Smart Device based on Scenario", Journal of The Korea Society of Computer and Information, Vol. 17, No. 11, 2012.

[3] Lee, Youja and Yu, Hongsik, "Effects of Broadcasting Content Curation on Video-On-Demand Use Traffic : Based on a Case of the Educational Broadcasting System", Information Society & Media, Vol. 18, No. 3, 2017,

[4] Jongho, Choi and Keebeak, Kim, "Development of real time clouding-contents curation platform based on hybrid analysis", Research Report, Korean Small and Medium Business Administration, 2017.

[5] David M W Powers, "Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness and Correlation", Journal of Machine Learning Technology, 2011.

[6] Perruchet, P., Peereman, R., "exploitation of distributional information in syllable processing", Journal of Neurolinguistics, 2004.

[7] David M W Powers, "The Problem with KappaEACL", Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics, 2012.

[8] Fawcett, T., "An Introduction to ROC Analysis", Pattern Recognition Letters. 2006.

[9] Dongwon Kim, Mi-Hee Youn, 'Performance Evaluation on the Power Consumption of IEEE802.15.4e TSCH', The Journal of The Institute of Internet, Broadcasting and Communication VOL. 18 No. 1, 2018

저자약력

최 종 호(Jong-Ho Choi) **[종신회원]**

- 1982년 2월 : 중앙대학교 전자공학과(공학사)
- 1984년 2월 : 중앙대학교 대학원 전자공학과(공학석사)
- 1987년 2월 : 중앙대학교 대학원 전자공학과(공학박사)
- 1990년 3월 ~ 현재 : 강남대학교 IoT전자공학과 교수

〈관심분야〉 영상처리, 컴퓨터시각, 딥러닝