Journal of Digital Convergence
Vol. 17. No. 11, pp. 201-206, 2019
ISSN 1738-1916
https://doi.org/10.14400/JDC.2019.17.11.201

협업필터링을 활용한 보험사 웹 사이트 내의 콘텐츠 추천 시스템 제안

강지영[1], 임희석[2*]
[1]고려대학교 컴퓨터정보통신대학원 석사과정, [2]고려대학교 컴퓨터학과 교수

Proposal of Content Recommend System on Insurance Company Web Site Using Collaborative Filtering

Jiyoung Kang[1], Heuiseok Lim[2*]
[1]Graduate School of Computer & Information Technology, Korea University, Master's Course
[2]Department of Computer Science and Engineering, Korea University, Professor

요 약 온라인에서 보험 정보를 찾는 이용자들이 많은 반면, 보험사 웹 사이트 콘텐츠 추천 연구 사례는 많지 않았으므로 본 연구에서는 보험사 웹 사이트의 페이지 방문 이력을 활용하여 사용자에게 선호 가능성이 높은 페이지 추천 시스템을 제안하였다. 데이터는 웹 브라우저 이용 시 발생하는 클라이언트 사이트 스토리지(Client-side storage)를 활용하여 수집하였으며, 추천 기술로는 협업 필터링(Collaborative filtering)을 연구에 적용하였다. 실험을 실시한 결과 방문 여부를 의미하는 이진화된 데이터를 사용한 자카드 인덱스(Jaccard index) 기반의 아이템 기반 협업 필터링(Item-based collaborative, IBCF)에서 좋은 성능을 나타내었다. 향후에는 아이템에 가중치를 부여한 추천 기술을 연구하여, 기업에서 사용 시 마케팅 전략에 부합하는 콘텐츠 추천 시스템을 구현할 수 있을 것이다.

주제어 : 추천 시스템, 협업 필터링, IBCF, 자카드 인덱스, 클라이언트 사이드 스토리지

Abstract While many users searched for insurance information online, there were not many cases of contents recommendation researches on insurance companies' websites. Therefore, this study proposed a page recommendation system with high possibility of preference to users by utilizing page visit history of insurance companies' websites. Data was collected by using client-side storage that occurs when using a web browser. Collaborative filtering was applied to research as a recommendation technique. As a result of experiment, we showed good performance in item-based collaborative (IBCF) based on Jaccard index using binary data which means visit or not. In the future, it will be possible to implement a content recommendation system that matches the marketing strategy when used in a company by studying recommendation technology that weights items.

Key Words : Recommendation system, Collaborative filtering, IBCF, Jaccard index, Client-side storage

1. 서론

추천 시스템은 상품을 추천하거나 예측하기 위해 사용자들의 상품 구매 정보를 분석하여, 다른 사용자들 중 비슷한 성향의 사용자들을 찾는다[1-2]. 추천 기술 중 협업 필터링(Collavorative filtering)은 유사한 사용자 또는 아이템에 대한 정보를 기반으로 한 알고리즘이며,[3] 개인 웹페이지 구성이나 개인 맞춤형 서비스가 필요한 광고 등

*Corresponding Author : Heuiseok Lim(limhseok@korea.ac.kr)
Received September 18, 2019 Revised October 21, 2019
Accepted November 20, 2019 Published November 28, 2019

다양한 분야에 적용이 가능하다[4].

상품 추천 알고리즘은 책, 음악, 여행 등과 같이 상품 구매가 이루어지는 쇼핑몰에서 주로 응용되고 있는데, 결제가 아닌 정보 탐색을 위해 제공한 보험사 웹사이트 내의 콘텐츠 추천 목적으로 추천 알고리즘이 적용된 사례는 알려지지 않았으므로, 본 논문에서는 보험사 웹 사이트의 방문 이력을 수집하여 그 특성에 맞는 추천 시스템을 제안하였다. 보험회사는 보험상품이 소비자에게 복잡하다[5]는 특성을 고려하여 웹을 방문하는 사용자들에 적합한 판매 상품을 제시하거나 정보 제공[6,7]을 위한 콘텐츠 추천을 통해 고객이 선호할 만한 웹페이지를 안내해 주는 전략이 필요하다.

본 논문에서는 1장 서론에 이어 2장에서는 추천 기술 소개와 유사도 측정 방법을 설명한다. 또한 데이터 수집을 위해 적용한 클라이언트 사이드 스토리지 저장 방식(Client-side storage)을 기술한다. 3장에서는 웹페이지 방문 이력 데이터에 협업 필터링 추천 기술을 적용하여 결과를 도출하고 성능 평가를 하였으며, 이를 통해 4장에서는 최적화된 추천 시스템을 제안하고 5장에서는 향후 연구 과제를 논의한다.

2. 관련 연구

2.1 추천 시스템

2.1.1 추천 기술의 종류

추천 시스템은 추천을 위해 검토 가능한 사용자의 디지털 흔적(digital footprint)이나 제품 관련 정보들을 모두 고려해 사용자에게만 개인화된 추천을 제공한다. 널리 사용되는 추천 시스템으로 협업 필터링 추천 시스템, 콘텐츠 기반 추천시스템, 지식 기반 추천 시스템, 하이브리드 시스템이 있다[8].

협업 필터링(Collaborative filtering)은 두 명의 사용자가 과거에 비슷한 관심사를 가지고 있다면, 미래에도 비슷한 취향을 가질 것이라는 아이디어로, 유사한 사용자 또는 아이템에 대한 정보를 기반으로 한 알고리즘이다[9]. 협업 필터링은 보통 두 가지 종류로 나눈다. 아이템 기반 협업 필터링(Item-based collaborative filtering, IBCF)은 사용자가 이전에 구매한 것들과 가장 유사한 아이템을 추천한다[10]. 사용자 기반 협업 필터링(User-based collaborative filtering, UBCF)은 유사한 사용자가 가장 선호하는 아이템을 사용자에게 추천한다. 본 연구에서는 방문 페이지 URL을 기록 하였으므로, 사용자 클릭에 따라 중복 이력이 발생할 수 있으며, 사용자가 평가를 남긴 것이 아니므로 사용자에 대한 정확한 정보가 있다고 할 수 없다. 이러한 제한 된 상황에서는 IBCF 기법 사용이 적합할 수 있다.

2.1.2 유사도 계산

협업 필터링의 기본 단계는 사용자 간 혹은 아이템 간 유사도를 계산하는 것이다. 협업 필터링에서 가장 널리 쓰이는 유사도 측정치에는 코사인(Cosine) 유사도가 있다. Fig. 1을 이용하여 UBCF의 코사인 유사도를 설명하면, 각 사용자의 평가 값을 좌표로 사용해서 원점에서 사용자까지의 선을 그었을 때, 각도가 작아질수록 코사인 유사도 값이 1에 가깝게 되며, 반대로 코사인 유사도 값이 0이라는 것은 아무런 관련성이 없음을 의미한다[11].

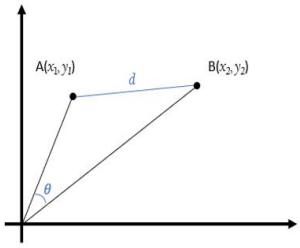

Fig. 1. Cosine similarity

만일 데이터가 이진 값을 갖는다면 코사인 유사도를 사용하지 않고, 자카드 인덱스(Jaccard Index)를 사용 하여 식 (1)와 같이 A와 B가 동시에 선택한 아이템 개수에서 A와 B중 하나라도 선택된 항목의 개수로 나눈다.

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|} \quad (1)$$

2.1.3 성능 평가

본 연구에서는 Table 1의 오분류표(Confusion Matrix)을 활용해 모델을 평가하기 위한 정확도/재현율/TPR/FPR 지표들을 계산한다.

Table 1. Confusion Matrix

		Actual	
		Positive	Negative
Predicted	Positive	True Positive	False Positive
	Negative	False Negative	True Negative

(1) 정확도(Precision): True로 예측한 관측지 중 실제값이 True인 것의 비율을 나타내는 정확성(exactness) 지표이다.
(2) 재현율 (Recall), TPR(True Positive Rate): 실제값이 True인 관측치 중 예측치가 적중한 정도를 나타내며 모형의 완전성(completeness)를 평가하는 지표이다.
(3) FPR(FP Ratio): 추천 시스템이 추천했는데 사용자가 선택하지 않은 아이템의 비율이다[12].

위 지표를 수식으로 표현하면 식 (2)과 같다.

$$precision = \frac{TP}{TP + FP}$$

$$recall = TPR = \frac{TP}{TP + FN}$$

$$FPR = \frac{FP}{FP + TN} \quad (2)$$

정확도(precision)와 재현율(recall)은 특히 텍스트를 분류하고 정보를 검색하는 경우에 자주 사용되며, 한 지표의 값을 개선하면 다른 지표의 값은 낮아지는 관계가 나타날 가능성이 높다.

ROC(Receiver Operator Characteristic) 커브는 분석 결과를 가시화할 수 있다는 점에서 유용한 평가 도구이며, 그래프의 밑부분 면적 AUC(Area Under the Curve)가 넓은 면적으로 계산될수록 좋은 모형으로 평가한다. Fig. 2의 경우 모델 A를 더 높은 성과를 가지는 것으로 판단한다.

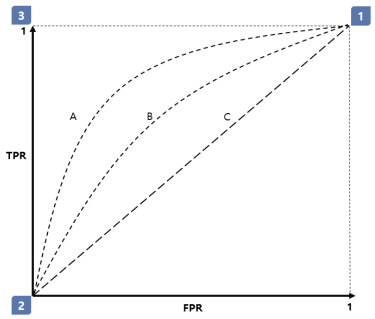

Fig. 2. ROC Curve

2.2 클라이언트 사이드 스토리지(Client-side storage)

웹사이트 운영자는 방문한 사용자의 디바이스에 소량의 정보를 기록하여 개인화된 서비스를 제공하거나[13] 사이트 기능을 개선하기 위한 목적으로 활용할 수 있다.

이러한 클라이언트 사이드 스토리지(Client-side storage) 기능에는 쿠키(cookie)가 있으며, HTML5에서는 쿠키의 단점을 보완하여 웹 스토리지(web storage) 기능을 만들었다. 이것은 쿠키 사이즈 4KB보다 큰 5MB까지 저장 가능하며 HTTP 요청마다 전달하지 않는다. 이는 쿠키에 비해 불필요한 네트워크 트래픽이 발생하지 않는 장점이 있다[14]. 웹 스토리지에는 로컬 스토리지(local storage)와 세션 스토리지(session storage)가 있다. 로컬 스토리지는 사이트 재방문 시 이전에 저장되었던 정보를 이용 할 수 있으므로 활용도가 높다. 세션 스토리지는 세션 종료 시 데이터가 만료되는 특징이 있다.

대부분의 웹 사이트에서는 이용자에게 설명 시 '로컬 스토리지'도 일반적인 의미로 '쿠키'라고 총칭하고 있으며, 웹 사이트 개인정보처리방침에서는 접속 파일의 운영에 관한 정보를 이용자에게 공개하고 이용자가 웹 쿠키를 원하지 않을 경우 거절할 수 있으며 이것이 서비스에 영향을 주어서는 안 된다고 명시한다[15].

3. 협업필터링을 이용한 웹 사이트 내의 콘텐츠 추천 모델

3.1 데이터 수집

Fig. 3은 웹 페이지 방문 이력 저장을 위한 처리 흐름도이다. 신규 사용자가 웹 사이트 방문 시 웹 사이트 서버는 사용자 아이디를 발급하여 웹 서버와 클라이언트의 로컬 스토리지에 저장한다. 사용자가 페이지를 이동하면서 변경되는 웹 페이지 정보는 웹 사이트 서버에 저장한다. 웹 페이지는 URL로 구분 가능하나 분석 효율을 위해 코드 형식으로 변환하여 사용하였다.

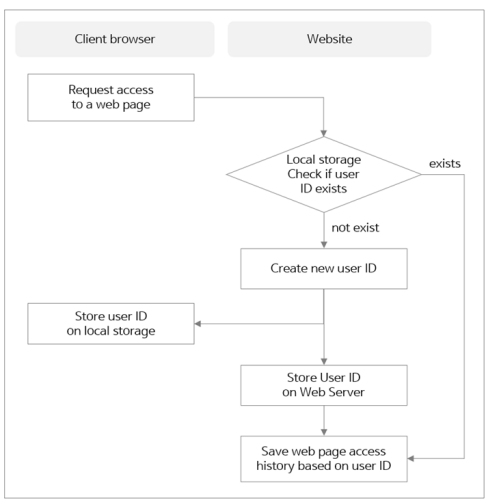

Fig. 3. Flowchart for saving web page history

3.2 데이터 전처리

총 7일간 수집된 웹 페이지 방문 이력 중에서 모델 적용에 적합하지 않은 사용자와 아이템을 제거하는 작업을 선행하였다.

사용자 아이디 중에서는 웹 페이지 2회 이하 방문한 사용자 아이디는 제외하였다. 해당 데이터에는 광고 배너를 실수로 클릭했거나, 사용자가 아닌 웹 크롤링에 의한 접속과 리디렉션(redirection)이 발생한 경우가 다수 포함되어 있다. 추천 아이템으로 모델에 적용할 웹 페이지 코드 중에서는 극단적으로 방문 횟수가 많은 웹 페이지와 매우 희소하게 방문 된 웹 페이지는 제외하였다. 극단적으로 많은 경우 콘텐츠 추천 대상으로 큰 의미 없는 메인 페이지인 경우가 대부분이다.

추천 모델에 따라 바이너리(binary) 데이터에서 작동하는 경우도 있으므로 방문 여부를 의미하는 0과 1로만 이루어진 이진화 된 데이터도 만들어 연구에 활용하였다. 웹 사이트 이용자 동선이나 웹 사이트의 리디렉션이 발생함에 따라 중복 로그가 발생할 수도 있으므로, 페이지를 몇 번 조회했는지 횟수로 계산하는 것 보다 방문을 했는지 안 했는지가 더 정확한 자료일 수 있다.

3.3 모델 구현

본 연구에서는 아래 4가지 모델을 사용하였다.

모델1) UBCF. 코사인유사도. 방문횟수 데이터

모델2) UBCF. 자카드인덱스. 방문여부 이진화 데이터

모델3) IBCF. 코사인유사도. 방문횟수 데이터,

모델4) IBCF. 자카드인덱스. 방문여부 이진화 데이터

각 모델에서 사용자 아이디에 따른 추천된 아이템 예시는 Table 2와 같다.

Table 2. Recommended Web Page Codes by User ID

User ID	Recommended web page code				
6737	10031	10002	10003	10004	10005
28923	10001	10002	10003	10004	10005
31274	10001	10002	10004	10005	10006
113873	10047	10041	10011	10042	10040
114697	10016	10014	10001	10002	10003

추천된 웹 페이지의 분포는 Fig. 4.과 Fig. 5.의 그래프를 통해 파악할 수 있다. x축은 웹 페이지 코드를 의미한다. 특정 웹 페이지 추천이 집중되었으나, IBCF 모델에서는 UBCF 모델에 비해 다양한 웹 페이지가 추천되었다.

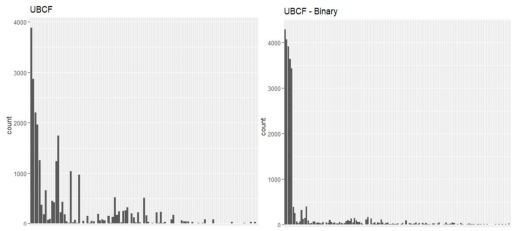

Fig. 4. UBCF. Distribution of recommended results using hit data(left) and distribution of recommended result using binary data(right)

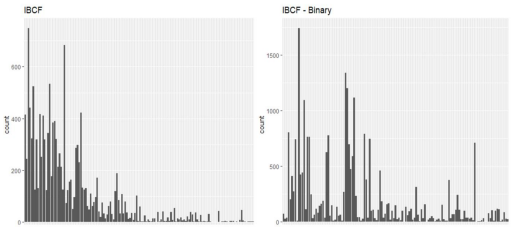

Fig. 5. IBCF. Distrubution of recommended results using hit data(left) and distribution of recommended result using binary data(right).

3.4 성능 평가

사용자 아이디별 추천 페이지 결과의 TPR, FPR ROC 곡선은 Fig. 6이며, 정확도별 재현력 계산 결과는 Fig. 7이다. 이 두 가지 그래프를 통해 자카드 인덱스 기반의 IBCF 모델을 선택하였다.

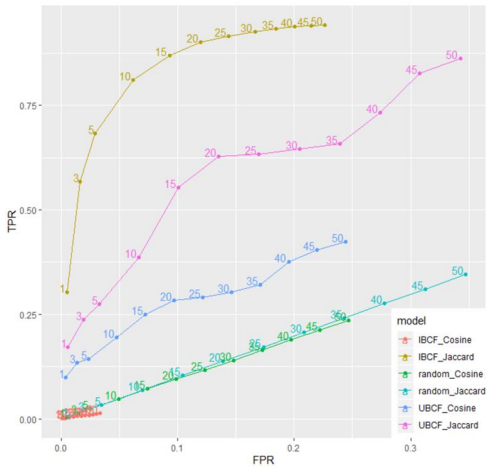

Fig. 6. ROC Curve

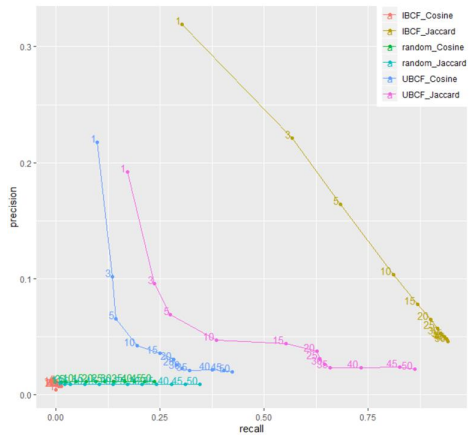

Fig. 7. prec/rec

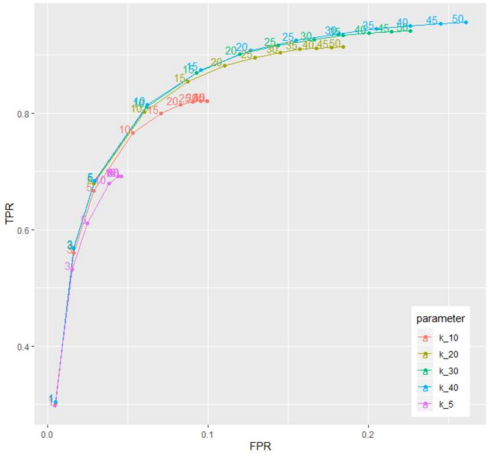

Fig. 8. ROC curve by parameter k

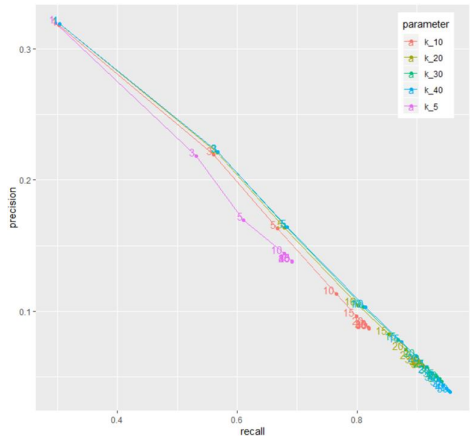

Fig. 9. prec/rec by parameter k

마지막으로 선택한 모델의 매개변수 최적화 과정을 수행하였다. IBCF는 가장 가까운 k개의 아이템을 고려하므로, 이 k를 최적화하기 위해 다양한 탐색을 해 볼 수 있다. 매개변수 k를 5~40 범위로 탐색한 성능지표 TPR, FPR는 Fig. 8이며, Fig. 9는 매개변수 k를 5~40 범위로 탐색한 정확도, 재현력 결과이다. 이 결과를 통해 k=40이 최적화된 매개변수라고 할 수 있다.

4. 연구 결과 및 향후 과제

본 연구에서 선택한 모델을 통해 웹 로그를 활용한 추천 시스템 구현에서는 방문 횟수보다 방문 여부만 사용하는 것이 효과적이며, 웹 사이트 방문자 간의 유사도 보다 웹 페이지 간의 유사도를 기반으로 기술을 적용하는 것이 타당함을 알 수 있다.

IBCF 모델은 가까운 k개의 아이템을 고려하므로 k를 최적화 할 수 있도록 매개변수 값을 변경하여 평가하였다. ROC curve와 prec/rec 그래프를 통해 k=40이 최적화된 매개변수로 결정되었다.

추천 시스템의 결과를 활용하면 웹 사이트를 방문한 고객에게 추천 콘텐츠를 제시하여 연관성이 높은 다른 페이지로 이동을 유도 할 수 있다. 인구통계학적 요인이나 구매 이력으로 광고나 콘텐츠를 제시하는 형태가 아닌, 관심을 보였던 콘텐츠, 페이지 방문 이력을 기반으로 개인 맞춤형 콘텐츠를 제시하는 형태라고 할 수 있다. 본 연구 외에 추가로 제안하는 것은 아이템의 특징을 이용한 추천 방법과 결합하여, 정보 과잉으로 선택이 어려운 상황에서 더 좋은 선택이 가능하도록 큐레이션(curation) 정보를 제공하는 것이다. 이해하기 어려운 정보가 많은 보험 웹 사이트에서는 더 자주 노출이 필요한 콘텐츠에는 가중치를 부여하여 마케팅 전략에 부합하는 아이템 추천 기술이 필요하다.

REFERENCES

[1] J. W. Kim & K. H. Park. (2016). Personalized Group Recommendation Using Collaborative Filtering and Frequent Pattern. *The Journal of Korean Institute of Communications and Information Sciences, 41(7),* 768-774.

DOI:10.7840/kics.2016.41.7.68

[2] S. H. Park, J. W. Kim, D. H. Kim & H. J. Cho. (2019). Music Therapy Counseling Recommendation Model Based on Collaborative Filtering. *Journal of the Korea*

Convergence Society, 10(9), 31-36,
DOI:10.15207/JKCS.2019.10.9.031

[3] S. K. Gorakala & M. Usuelli. (2015). *Building a Recommendation System with R,* Birmingham: Packt Publishing.

[4] I. Lim. (2016). *Recommendation system using R.* Seoul: Chaos Book.

[5] H. J. Sim, M. J. Kim & H. C. Choe. (2018). Consumer's Satisfaction of Insurance Consumption : Focusing on Self-determination Theory. *Journal of the Korea Convergence Society, 9(5),* 157-169,
DOI:10.15207/JKCS.2018.9.5.157

[6] M. Field & V. Stoykov. (2007. June). Online branding: the new frontier[online]. *InFinance: The Magazine for Finsia Members, 121(2).*

[7] J. H. Park. (2014. Oct). Online Channel Activation Plan by Diversifying Sales Channels. *KIRI(Korea Insurance Research Institute) Weekly, 305,* 1-4.

[8] E. Y. Bae & S. J. Yu. (2018). Keyword-based Recommender System Dataset Construction and Analysis. *Journal of KIIT, 16(6),* 91-99.
DOI:10.14801/jkiit.2018.16.6.91

[9] S. K. Gorakala & M. Usuelli. (2015). *Building a Recommendation System with R,* Birmingham: Packt Publishing.

[10] J. W. Choi. (2018). *A Study for Improving Sparsity and Scalability Problem in Collaborative Filtering Recommendation System.* Master dissertation. Soongsil University, Seoul.

[11] S. R. Jung. (2018). A Study on Improving Efficiency of Recommendation System Using RFM. *Journal of the Korean Institute of Plant Engineering, 23(4),* 57-64.

[12] Korea Data Agency. (2019). *The Guide for Advanced Data Analytics Professional.* Seoul: Korea Data Agency

[13] Mozilla contributors. (n.d.). *HTTP cookies*, MDN Web Docs. *https://developer.mozilla.org*

[14] H. W. Myeong. J. H. Paik & D. H. Lee. (2012). Study on implementation of Secure HTML5 Local Storage. *Journal of Korean Society for Internet Information, 13(4),* 83-93. DOI:10.7472/jksii.2012.13.4.83

[15] Y. H. Kim & T. S. Lee. (2014. Aug). Analysis of Major Issues and Vulnerabilities in Internet Cookies. *Internet & Security Focus,* 79-98.

강 지 영(Jiyoung, Kang) [학생회원]

· 2005년 8월 : 동국대학교 컴퓨터공학과(공학학사)
· 2015년 9월 ~ 현재 : 고려대학교 컴퓨터정보통신대학원 석사과정
· 관심분야 : 추천시스템, 마케팅
· E-Mail : kangjy1002@gmail.com

임 희 석(Heuiseok, Lim) [종신회원]

· 1992년 2월 : 고려대학교 컴퓨터학과(이학학사)
· 1994년 2월 : 고려대학교 컴퓨터학과(이학석사)
· 1997년 8월 : 고려대학교 컴퓨터학과(이학박사)
· 2008년 3월 ~ 현재 : 고려대학교 컴퓨터학과 교수
· 관심분야 : 자연어처리, 뇌신경 언어 정보 처리
· E-Mail : limhseok@korea.ac.kr