The Journal of Digital Policy & Management 2013 Jul; 11(7): 193-199

사용자 검색이력 기반의 잠재적 질의어 추천 시스템 개발*

박정배*, 박기남**, 임희석*
고려대학교 컴퓨터교육과*, 순천향대학교 컴퓨터소프트웨어공학과**

Development of the Potential Query Recommendation System using User's Search History

Jeongbae Park*, Kinam Park, Heuiseok Lim***
Dept. of Computer Science Education, Korea University*
Dept. of Computer Software Engineering, Soonchunhyang University**

요 약 본 논문에서는 정보검색 시스템 사용자가 자신의 잠재적 정보욕구를 질의어로 표현하고, 원하는 정보가 검색될 수 있도록 사용자 검색이력 기반의 잠재적 질의어 추천 시스템을 제안한다. 제안하는 시스템은 사용자의 검색 질의어를 기반으로 기존 사용자들의 검색이력과의 연관관계를 분석하고, 사용자 잠재적 정보욕구를 추출하였다. 추출된 잠재적 정보욕구는 추천 질의어로 표현되어 사용자에게 추천된다. 본 논문에서는 제안한 시스템의 효용성 분석을 위하여 27,656건의 검색이력 데이터를 이용하여 행동실험을 실시하였다. 실험결과 피험자들은 제안한 시스템을 사용할 때 일반 검색엔진을 사용할 때 보다 높은 통계적으로 유의미한 만족도를 나타내었다.

주제어 : 질의어추천, 정보검색, 연관규칙, 검색이력, 잠재적 질의어

Abstract In this paper, a user search history based potential query recommendation system is proposed to enable the user of information search system to represent one's potential desire for information in terms of query and to facilitate the desired information to be searched. The proposed system has analyzed the association with the existing users's search histories based on the users' search query, and it has extracted the users's potential desire for information. The extracted potential desire for information is represented in terms of recommended query and thereby made recommendations to users. In order to analyze the effectiveness of the system proposed in this paper, we conducted behavioral experiments by using search histories of 27656. As a result of behavioral experiments, the experiment subjects were found to show a statistically significant higher level of satisfaction when using the proposed system as compared to using general search engines.

Key Words : Query recommendation, Information retrieval, Relevant rule, Search history, Potential query

* 본 논문은 2012년도 산학협동재단의 학술연구과제 및 순천향대학교의 교내연구지원과제에 의해 수행됨
Received 11 June 2013, Revised 15 July 2013
Accepted 20 July 2013
Corresponding Author: Heuiseok Lim(Korea University)
Email: limhseok@korea.ac.kr

ISSN: 1738-1916

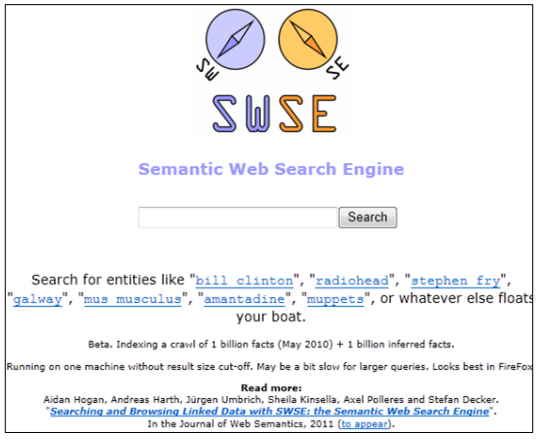
1. 서론

정보검색이란 사용자의 정보요구에 따라 기존 데이터를 분석한 후 사용자 정보요구에 적합한 정보를 탐색하는 일련의 과정을 의미한다. 이러한 과정 중 사용자가 정보욕구를 표현하기 위한 질의어는 검색결과에 많은 영향을 미치게 된다[1]. 현재 상용화된 검색엔진들은 사용자가 해당분야에 대한 충분한 이해와 사전지식이 없으면 원하는 정보를 탐색하기 힘들며, 정보에 대한 핵심키워드를 찾고, 이를 질의어로 표현해야 하는 부가적인 노력이 필요하다[1, 2]. 따라서 사용자들은 정보검색에 적합한 질의어를 찾는데 많은 시간을 허비하게 되고, 적합한 질의어를 찾지 못한 사용자들은 반복된 검색과정을 통해 질의어를 수정해 나간다. 최악의 경우 사용자는 원하는 정보를 탐색하지 못하고 정보검색을 포기하게 된다.

일반적으로 검색에 사용되는 질의어는 길이가 짧고, 의미가 중의적이며 함축적인 경우가 많고, 시스템 측면에서도 질의어 길이가 3개 단어를 초과하면 검색 적합성이 낮아지는 경향이 있다[3, 4]. 이와 같은 문제점을 해결하기 위한 대표적인 연구방법으로 질의어확장(query expansion)과 연관 검색어(relevant keyword)가 있다[5]. 질의어확장은 정보검색 시스템이 사용자 질의어에 자동으로 부가적인 정보를 추가하여 검색을 수행하는 방법이다[1]. 부가정인 정보란 지식기반(knowledge-based), 통계기반(statistical-based) 그리고 개념기반(concept-based) 등으로 추출된 정보를 의미한다. 지식기반은 일반적으로 시소러스(thesaurus)를 활용하여 부가적인 정보를 추출한다. 하지만 시소러스는 구축하기가 용이하지 않으며, 단어의 희귀(sparseness) 문제를 극복하기 어려운 단점이 있으며, 단어의 중의성으로 인해 정확률이 떨어지는 문제점이 있다[3]. 통계기반은 질의어를 구성하는 단어의 동시 출현 횟수를 이용하는 방법으로써 출현하는 단어들이 같은 주제에 밀접하게 관련되어 있음을 가정한다. 하지만, 동시 출현 빈도가 높은 단어들은 연관문서 뿐만 아니라 비연관문서의 대표단어로 나타날 수 있기 때문에 또 다른 문제점이 나타날 수 있다[3]. 마지막으로 개념기반은 검색에 활용될 전체 문서에서 출현빈도를 고려한 단어를 추출하여 구축한 의미망을 이용한다. 하지만 용어간의 개념관계를 설정하기 어렵다. 연관검색어는 사용자의 질의어를 분석하여 기존 사용자들의 질의어 이력에서 유사한 검색 질의어들을 선정하여 사용자에게 추천하는 방법이다. 하지만 이 방법은 주어진 질의어를 구성하고 있는 단어와 가장 밀접한 단어를 선택할 때 사용빈도에만 의존할 수밖에 없다[1]. 이에 본 논문에서는 정보검색 시스템 사용자가 자신의 잠재적 정보욕구를 질의어로 표현하고, 원하는 정보가 검색될 수 있도록 사용자 검색이력 기반의 잠재적 질의어 추천 시스템을 제안한다.

2. 관련 연구

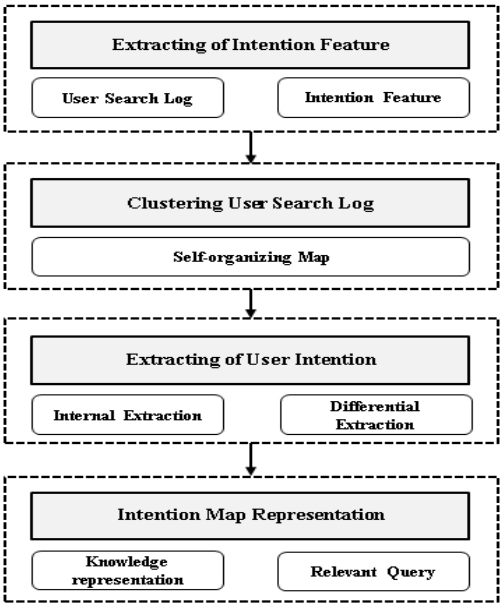
정보검색 시스템 사용자들은 질의어로 정보욕구를 표현할 뿐만 아니라, 미처 질의어로 표현하지 못하는 잠재적 정보욕구를 갖고 있기 때문에 시스템이 이를 자동으로 파악하기란 쉽지 않다. 또, 질의어 특성상 길이가 짧고, 중의적 의미를 내포하는 단어가 많기 때문에 시스템이 사용자 정보욕구를 예측하기에는 정보량이 충분지 않다[1]. 이러한 문제점을 해결하기 위해 검색 알고리즘과 정보의 신뢰도를 강화하려는 전통적인 검색 시스템 연구 이외에도 사용자의 검색 욕구를 분석하기 위한 연구가 활발히 진행 중이다.



[Fig. 1] SWSE(Semantic web search engine)

[Fig. 1]은 DERI 연구소에서 개발한 SWSE 검색엔진으로 시맨틱 웹 표준 언어로 표현되는 정보를 객체 지향적인 관점에서 검색하거나, 네비게이션 해주는 검색엔진으로 최초 검색어로부터 정보를 객체단위로 찾아가는 인

터페이스를 지원한다[6]. 웹상의 텍스트 문서를 찾는 개념이 아닌 객체 단위로 RDF 자원을 검색하는 것으로 범위를 넓혔다. SWSE는 약 10억 건의 RDF 문서를 URI를 구분하여 수집하고 검색 서비스를 제공하고 있다. SWSE는 내부 쿼리 엔진으로 W3C 표준 질의 언어인 SPARQL을 사용하여 질의를 처리한다.



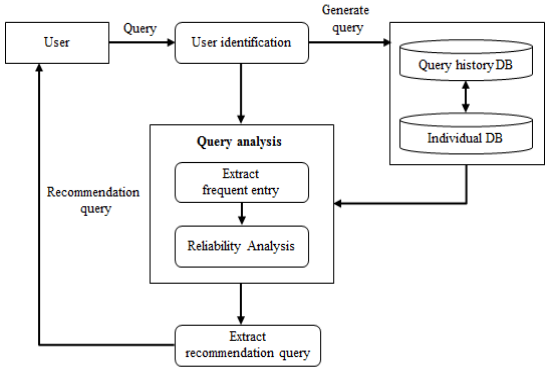
[Fig. 2] Automatic extraction of user's search intention from web search logs

[Fig. 2]는 사용자 검색 이력 데이터를 이용하여 사용자 의도 추출 방법을 제안하였다. 검색 성능 향상을 위해 기존 사용자들의 검색이력 데이터를 바탕으로 사용자 검색 의도 자질을 선정하고, 클러스터링 알고리즘과 사용자 검색 의도 추출 알고리즘을 이용하여 사용자 검색 의도를 자동으로 추출하였다. 모델은 사용자를 고려하지 않은 인터페이스를 제공하고 있으며, 빈도 기반의 검색 알고리즘을 사용함으로써 검색 결과 품질의 신뢰성을 확보하지 못하는 단점이 있다[1].

[7]은 웹 기반 정보검색에서 개인화 검색 측면의 사용자의도의 중요성을 강조하면서 서로 다른 의도를 갖는 개인 사용자의 정보 욕구를 충족 시켜줄 수 있는 HPVM(Hierarchical Phrase Vector Mode)을 제안하였다. HPVM은 구 (phrase) 기반의 벡터모델을 통해 사용자 Intention을 계층적으로 확장하고, SVM(Support Vector Machine)을 이용해 학습한 사용자 Intention을 토대로 부정적(positive) 이거나 부정적(negative) 문서 분류를 통해 가장 사용자 의도에 적합한 문서를 사용자에게 제공한다. 하지만HPVM은 사용자 피드백을 통해 질의어를 확장하고, 패턴을 고려하였기 때문에 사용자 참여가 이루어지지 않을 경우 성능에 영향을 줄 수 있다. 또한 초기 질의어 패턴 학습에 대한 내용이 고려되지 못했다.

3. 잠재적 질의어 추천 시스템

본 논문에서는 기존 연구들의 사용자 정보욕구 및 검색의도 파악 미흡, 중복적인 사용자 정보욕구에 따른 비용 증가 등의 문제점을 해소하기 위해 사용자 정보이력 기반의 잠재적 질의어 추천 시스템을 제안한다. 제안하는 시스템은 사용자의 검색이력을 이용하여 사용자 잠재적 정보욕구 후보를 추출하고, 추출된 후보와 기존 사용자들의 검색이력 간의 연관관계 판단하여 잠재적 정보욕구를 추출하였다. 추출된 잠재적 정보욕구는 질의어로 표현되어 사용자에게 추천된다. [Fig. 3]은 본 논문에서 제안하는 사용자 검색이력 기반의 잠재적 질의어 추천 시스템 구성도이다.



[Fig. 3] The system architecture

제안 시스템에서는 사용자의 잠재적 질의어 추천을 위해 사용자가 입력한 질의어를 바탕으로 개인이력과 대중이력을 기반으로 질의 분석 및 추천 질의어를 추출한

다. 개인이력기반으로 빈발항목 추출 후 신뢰도를 계산하고, 대중이력을 바탕으로 다시 빈발항목 추출 및 신뢰도를 측정한다. 즉, 개인이력기반 연관규칙탐색 신뢰도가 최소신뢰도를 만족하지 못하는 경우 해당 질의어에 대한 개인이력은 잠재적 추천 질의어가 될 수 없으므로 유사한 관심을 표방한 대중이력 기반 빈발항목 추출 및 신뢰도 추출 후 최소 신뢰도 이상을 추천 질의어로 제시한다. 개인이력기반과 대중이력기반 모두 최소신뢰도 이상의 결과가 나오면, 개인이력기반, 대중이력기반 순서로 질의어를 추천한다.

3.1 사용자 식별

사용자 식별 처리는 개인이력을 대중이력과 분리하기 위한 방법이다. 이를 위해 제안 시스템에서는 로그인 방식과 쿠기 방식을 사용하였다. 로그인 방식은 시스템 환경에 제약을 받지 않는 장점이 있고, 쿠기 저장 방식은 장기적인 데이터수집에 장점이 있다. 시스템에서는 사용자가 질의어를 입력 시 데이터베이스에 질의어를 시계열로 저장한다. 이때 사용자 식별을 위해 고유 식별자 킷값을 함께 저장하며, 사용자 초기질의어와 동일 질의어 출현 시 가중치를 부여한다.

3.2 질의어 분석 및 질의어 추천

본 논문에서는 질의어이력 데이터베이스에 저장된 검색이력 중 추천질의어 후보 추출을 위해 사용자 입력질의어와 질의어이력 간의 연관관계를 분석하였다. 연관관계 분석은 개인이력 기반 신뢰도 계산과 대중이력 신뢰도 계산을 통해 이루어지며, 신뢰도 계산을 위한 빈발항목은 일자별 검색이력 단위로 추출하여 사용하였다. 본 논문에서는 연관규칙 분석을 위해 Apriori 알고리즘을 사용하였다[2]. Apriori 알고리즘은 비교 항목의 수가 늘어남에 따라 계산량이 증가하는 단점이 있긴 하나, 계산방법이 간단하고, 결과 이해가 용이하다는 장점이 있다. 그래서 본 논문에서는 일일 배치작업을 통해 연관관계 분석을 극대화 하였다.

질의어 분석 및 질의어 추천 과정을 예시를 통해 설명하면 다음과 같다. <Table 1>은 사용자 “A”, “B”, “C”, “D”, “E” 의 검색로그를 나타내고 있으며, 사용자별로 최초 입력질의어와 재입력한 질의어로 구성된다.

〈Table 1〉 Query logs

User	Query
A	{mango, onion, nintendo, milk, egg, yoplait}
B	{doll, onion, nintendo, milk, egg, yoplait}
C	{mango, apple, milk, egg}
D	{mango, umbrella, corn-flakes, milk, yoplait}
E	{corn-flakes, onion, onion, milk, ice-cream, egg}

사용자 입력질의어에 대한 로그는 빈발 항목집합 추출을 위하여 사용자별로 동일질의어 빈도가 계산된다. 단, 동일 사용자의 중복된 동일질의어는 하나의 트랜잭션으로 처리한다. <Table 2>는 <Table 1>을 이용한 빈발항목 후보 집합 추출 결과를 나타낸다.

〈Table 2〉 Frequent entry candidates

Query	Frequency
mango	3
onion	3
milk	5
egg	4
yoplait	3

<Table 2>에 나타난 후보항목 집합 군을 다시 검색로그 트랜잭션과 비교하여 빈발 항목집합을 찾아내는 과정을 반복하게 된다. <Table 3>은 검색로그 이력에 대한 빈발항목 집합을 나타낸다.

〈Table 3〉 Frequent entry

Query	Frequency
mango → onion	1
mango → milk	3
mango → egg	2
mango → yoplait	2
onion → milk	3
onion → egg	3
onion → yoplait	2
milk → egg	4
milk → yoplait	3
egg → yoplait	2

<Table 4>는 검색로그 이력에 대한 빈발항목 집합에서 임의로 설정한 가중치 이상의 항목을 추출한 내용이다. 본 논문에서는 최소 가중치를 3으로 설정하였으며, 가중치 미만의 값을 제외하였다.

〈Table 4〉 Weighted of frequent entry

Query	Frequency
mango → milk	3
onion → milk	3
onion → egg	3
milk → egg	4
milk → yoplait	3

<Table 4>로부터 구한 최종 빈발항목집합은 <Table 5>와 같으며{양파, 우유, 달걀} 와 {우유, 달걀, 요플레}는 각각 3, 2 의 신뢰도를 보여주고 있으며 초기질의어 "양파"에 대해서"우유, 달걀", 또는 초기질의어 "우유" 에 대해서 "달걀, 요플레"를 제시할 수 있다.

〈Table 5〉 Final frequent entry

Query	Frequency
onion milk egg	3
milk egg yoplait	2

4. 실험 및 평가

본 논문에서는 제안한 시스템의 효용성 평가를 위해 행동 실험을 실시하였고, 분석을 통해 잠재적 질의어 추출 방법을 검증하였다. 행동실험은 사용자의 시스템 만족도를 측정하였다. 만족도 측정은 정보욕구가 있는 사용자에게 추천 질의어를 제공함으로써 검색에 대한 만족도 여부를 측정한 것이다.

4.1 실험 방법

실험은 피험자 30명을 대상으로 실시하였다. 실험용 시스템 구축을 위해 사용한 데이터는 2012년11월20일 부터 2013년 3월30일까지 수집한 피험자 30명의 검색이력로그와 상용화된 검색시스템의 검색로그 27,656건을 활용하였다. 실험방법은 동일한 주제에 대하여 개인이력기반과 대중이력기반 추천 질의어를 검색욕구가 있는 사용자에게 추천함으로 그에 대한 만족도를 측정하였다.

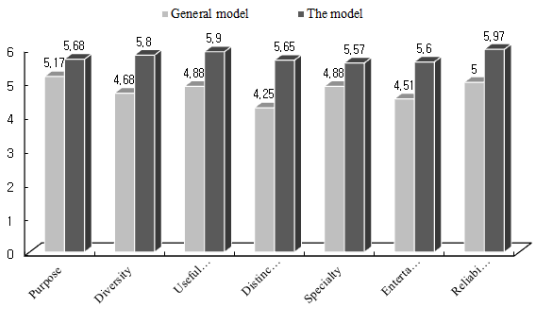
〈Table 6〉 Evaluation item

Item	Satisfaction Measurement
Purpose	Is appropriate information for query provided?
Diversity	Are the answers for the queries rich and various?
Usefulness	Are the answers for the queries complete and helpful?
Distinction	Are the answers for the queries distinguishable from the conventional search systems?
Specialty	Are the answers for the queries valuable above the level of common sense?
Entertainment	How much are you interested in the answers for the queries?
Reliability	Are the answers for the queries realistic and reliable?

만족도는 Likert Scale로 측정하였다. Likert Scale은 1932년 R.리커트가 고안한 태도 측정법으로 같은 종류의 내용에 관계되는 여러 의견을 모아 이것을 3~7단계의 연속체 척도상의 점수에 맞추어 그 합계 점을 가지고 태도의 점수로 삼는 상가평정척도와 항목분석에 의하여 작성하는 내적 일관성 척도를 결합시킨 태도 측정법이다. 만족도는 '매우 부적절함'을 1점, '매우 적절함'을 7점으로 하여 1점 단위로 체크할 수 있도록 구성하였다. 실험에 사용한 만족도 평가항목은 <Table 6>과 같다.

4.2 실험결과

본 논문에서는 잠재적 질의어 추천 만족도 검증을 위해 제안한 시스템과 상용화된 검색시스템의 만족도 에 대하여 paired T-test를 실시했다. 검증 결과 평균만족도는 제안한 시스템이 5.73("만족", SD=.31), 상용화된 시스템이 4.76("보통", SD=.15)으로 나타났다. 실험 참가자들은 제안한 시스템에 대해 더욱 만족하였으며, 이 차이는 $p<=0.5$에서 통계적으로 유의하게 나타났다 ($t(12) = -7.392$, $p<= 0.5$) [Fig.4].



[Fig. 4] Evaluate of user satisfaction

5. 결론 및 향후 과제

본 논문에서는 정보검색 시스템 사용자가 자신의 잠재적 정보욕구를 질의어로 표현하고, 질의어생성의 어려움 감소를 통해 원하는 정보가 검색될 수 있도록 사용자 검색이력 기반의 잠재적 질의어 추천 시스템을 제안하였다. 본 논문에서는 제안한 시스템의 효용성 평가를 위해 행동 실험을 실시하였고, 분석을 통해 잠재적 질의어 추출 방법을 검증하였다. 실험결과 피험자들은 제안한 시스템을 사용할 때 일반 검색엔진을 사용할 때 보다 높은 통계적으로 유의미한 만족도를 나타내었다.

본 논문에서 제안한 사용자 검색이력 기반의 잠재적 질의어 추천 연구의 의의는 다음과 같다. 먼저, 사용자 검색의도 자동 추출이 가능하다. 사용자 질의어에서 나타나는 키워드 추출 방법이 아닌, 잠재적 정보욕구를 사용자 검색이력에 기반을 두어 자동으로 추출할 수 있다. 또한, 사용자가 검색을 위한 질의어를 생성할 때 질의어 가이드를 제공함으로써 초기 검색 만족도를 향상 시킬 수 있다. 향후에는 개인화연구를 통해 검색이력을 확장 및 견고하게 함으로써 추천질의어에 대한 신뢰성을 확보할 수 있을 것이다.

ACKNOWLEDGMENTS

This research was supported by Korea Sanhak Foundation fund, 2012, and was partly supported by Soonchunhyang University, 2013.

REFERENCES

[1] Kinam Park, Hyesung Jee, Taemin Lee, Soonyoung Jung and Heuiseok Lim, Automatic extraction of user's search intention from web search logs, Multimedia Tools and Applications, Vol. 61, No. 1, pp. 145-162, 2012.

[2] Young-an Kim, Gun-Woo Park, An Efficient Extended Query Suggestion System Using the Analysis of Users' Query Patterns, Journal of the Korean institute of communication sciences, Vol. 37, No. 7, pp. 619-626, 2012.

[3] Ji-Hye Kim, Doo-Soon Park, Development of the Goods Recommendation System using Association Rules and Collaborating Filtering, The Journal of Korean association of computer education, Vol. 9, No. 1, pp. 71-80, 2006.

[4] Schuemie MJ, Kang N, Hekkelman ML, Kors JA, GeneE: gene and protein query expansion with disambiguation, Bioinformatics, Vol. 26, No. 1, pp.147-148. 2010.

[5] Xu, J., Croft, W. B. Query Expansion Using Local and Global Document Analysis. Proc. ACM SIGIR Int'l Conf. Research and Development in Information Retrieval, pp. 4-11, 1996.

[6] Andreas Harth, Aidan Hogan, Jürgen Umbrich, Stefan Decker, Building a Semantic Web Search Engine: Challenges and Solutions, Proceedings of the 3rd XTech Conference, 2008

[7] GunWoo Park, JinGi Chae, Dae Hee Lee, SangHoon Lee, User Intention based Personalized Search : HPS(Hierarchical Phrase Serch), the WSEAS International Conference on Applied Computing Conference, pp. 205-210, 2008.

박 정 배(Jeongbae Park)

- 2011년 2월 : 백석대학교 컴퓨터학과(공학사)
- 2013년 2월~현재 : 이니텍 차장
- 2012년 9월~현재 : 고려대학교 컴퓨터교육과 석사과정
- 관심분야 : 정보검색, 자연어처리
- E-Mail : insmile@korea.ac.kr

박 기 남(Kinam Park)

- 2004년 2월 : 백석대학교 컴퓨터학과 (공학사)
- 2006년 2월 : 한신대학교 컴퓨터정보학과(이학석사)
- 2011년 8월 : 고려대학교 컴퓨터교육학과(이학박사)
- 2013년 3월~현재 : 순천향대학교 컴퓨터소프트웨어공학과 교수
- 관심분야 : 인지과학, 정보검색, 자연어처리
- E-Mail : spknn@sch.ac.kr

임 희 석(Heuiseok Lim)

- 1992년 2월 : 고려대학교 컴퓨터학과 (이학학사)
- 1994년 2월 : 고려대학교 컴퓨터학과(이학석사)
- 1997년 2월 : 고려대학교 컴퓨터학과(이학박사)
- 2008년 3월~현재 : 고려대학교 컴퓨터교육과 교수
- 관심분야 : 컴퓨터교육, 자연어처리, 정보검색
- E-Mail: limhseok@korea.ac.kr