완전그래프를 이용한 문서요약 연구†

(Document Summarization Method using Complete Graph)

유 준 현*, 박 순 철**
(Jun Hyun Lyu · Soon Cheol Park)

요 약 본 논문에서는 웹 검색엔진에서 일반적으로 사용하는 문서요약에 대한 연구로써 문서내에 있는 문장들의 꼭짓점을 연결하는 완전그래프기법을 도입하여 요약내용을 좀 더 간결하고 함축하게 하는 통계요약기법을 제안했다. 이 요약기술을 지금까지 통계 문서요약기술에서 우수하다고 판단된 클러스터링 기법과 MMR 기법 등과 비교하였다. 특히, 요약 성능을 평가하기 위하여 인위적으로 요약된 요약문을 기준으로 한 각 요약기법들의 FScore값들과 비교하였다. 이 기술들 중에서 완전그래프기법이 약 30%정도 성능향상을 보였다.

핵심주제어 : 문서검색, 요약, 클러스터링, MMR, 완전그래프

Abstract In this paper, we present the document summarizers which are simpler and more condense than the existing ones generally used in the web search engines. This method is a statistic-based summarization method using the concept of the complete graph. We suppose that each sentence as a vertex and the similarity between two sentences as a link of the graph. We compare this summarizer with those of Clustering and MMR techniques which are well-known as the good summarization methods. For the comparison, we use FScore using the summarization results generated by human subjects. Our experimental results verify the accuracy of this method, being about 30% better than the others.

Key Words : Document Search, Summarization, Clustering, MMR, Perfect Graph

1. 서 론

인터넷의 사용자의 증가와 전자문서의 증가에 따라서 사용자들이 접근할 수 있는 문서의 양이 증가하고 있다. 수많은 문서 중 자신이 원하는 것을 찾기 위해서는 문서의 전체 내용을 읽어야 하지만 그것은 쉬운 일이 아니다. 이러한 문제점을 해결할 수 있는 방법이 문서요약기술이다.

문서요약의 경우 대표적으로 통계기반의 요약과 의미기반의 요약방법이 있다[1-4]. 의미기반 요약의 경우, 요약의 내용이 함축적이고 의미의 전달이 탁월할 수 있다는 장점은 있어 많은 연구가 진행 중이지만, 아직은 구현이 쉽지 않고 요약에 대한 신뢰성이 떨어진다[5].

통계기반의 경우, 일반적으로 정확성이 떨어지고 선택된 문장이 전체의 내용을 대표하지 못하는 단점이 있다. 그러나 구현이 용이해서 대부분 상용 문서요약시스템은 이 방법을 사용하고 있다.

본 논문에서는 단순 통계기반 문서요약의 단점을 보완하기 위하여 문서내의 문장을 꼭짓점으로 하는 완전그래프 기법을 이용하였다. 이 방법은 요약할 문서에 있는 각 문장과 모든 문서의 문장과의 관계를 모두 고려하였다. 또한 요약으로 추출되는 문장은 문서의 내용을 많이 포함하면서

† 이 연구는 전북대학교 2003년도 연구기반조성기금으로 이루어짐.
* LG 전자
** 전북대학교 전자정보공학부

다른 문장과는 유사성이 없는 문장을 선택한다.

이 방법의 성능 평가를 위해서 인위적 요약을 선행하여 비교하였다. 또한 중복내용 제거가 우수한 MMR(Maximal Marginal Relevance)기법[6-8]과 클러스터링기법[9,10]을 이용한 문서 요약을 본 연구와 비교하여 완전그래프 기법의 우수성을 보였다.

본 논문은 2장에서 본 논문에서 사용하고 있는 요약방법 소개하고, 3장에서 실험비교를 위하여 사용된 인위적 요약 결과를 설명하겠다. 4장에서 FScore 비교에 의한 실험결과를, 그리고 마지막으로 5장에서 본 논문의 결론을 이야기하겠다.

2. 요약방법

본 논문에서 사용하는 요약방법은 통계기반의 요약방법이다. 즉 문서 내의 문장과 용어의 통계값을 이용하여 중요 문장을 선택하는 방법이다. 특히, 본 논문에서는 요약된 문장의 중복성을 제거함으로써 기존의 방법보다 더 함축적인 요약내용을 구현하고자 한다.

2.1 MMR을 이용한 요약

MMR 알고리즘은 정보검색시스템의 검색결과에서 중복내용을 제거하여 결과를 보여주는데 사용되고 있다[6-8]. 본 논문에서는 이러한 MMR의 중복성 제거 특성을 이용하여 중복이 적고 많은 정보를 포함하는 것을 요약 문장으로 선택하였다.

문서내의 문장에서 사용되는 단어의 가중치의 계산 방법은 식 (1)과 같다. 단어의 가중치, s_{ij}는 j 문장에서 i번째 있는 단어, w_{ij}의 가중치이다.

$$s_{ij} = f_{ij} \cdot idf(w_{ij}) \cdot P(w_{ij}) \tag{1}$$

여기서,

$$f_{ij} = \frac{freq_{ij}}{freq_{ij}+2},\ idf(w_{ij}) = \log\frac{N}{n_{ij}},\ P(w_{ij}) = \begin{pmatrix} 2.0 & high \\ 1.5 & important \\ 1 & others \end{pmatrix}$$

이다.

식 (1)에서 $freg_{ij}$는 문서에서 단어 w_{ij}의 빈도수이며 $freg_{ij}$는 요약할 문서 내에 있는 단어 중 최대빈도수를 갖는 단어 w_{ij}의 빈도수이다. 따라서 f_{ij}는 $freg_{ij}$를 정규화한 단어 w_{ij}빈도의 정도를 나타내며 0부터 1사이의 실수값이다. 역문서빈도수 $idf(w_{ij})$의 계산식 중 N은 본 시스템에서 문서의 총 수이다. n_{ij}는 단어 w_{ij}가 출현한 문서의 수이다.

각 문서 내의 문장을 문장 벡터로 표현할 때 식 (2)와 같이 표현할 수 있다. 이 문장 벡터는 문장의 중요도를 계산할 때와 문장과 문장사이의 유사도를 계산할 때 사용되어진다[10].

$$\vec{S_i} = (s_{1,i}, s_{2,i}, \cdots, s_{n,i}) \tag{2}$$

문장의 중요도는 문장 벡터 내에 있는 단어들의 평균값으로 표현된다. 식 (3)은 문장의 중요도이다.

$$|\vec{S_j}| = \frac{\sum_{i=1}^{size\ of\ \vec{S_j}} s_{i,j}}{size\ of\ \vec{S_j}} \tag{3}$$

기본적인 MMR 통계요약 알고리즘은 식 (4)와 같다. MMR을 적용하지 않았을 때는 가중치가 높은 단어를 포함하는 문장을 중복하여 선택하게 된다. 이것을 보완한 것이 본 논문에서 사용하는 MMR[6-8] 기법이다.

$$arg \max_{S_i \in R-A} \{(|\vec{S_i}| - \lambda \cdot \max_{S_j \in A} \cdot (sim(\vec{S_i}, \vec{S_j}))\} \tag{4}$$

여기서 A는 요약 문장으로 선정된 문장의 집합이다. R은 문장의 중요도에 의해 정렬된 리스트이다. 유사도함수, $sim(S_i, S_j)$는 코사인 유사도를 사용하여 선택 가능한 문장과 이미 추출된 요약 문장 집합, A 에 포함된 문장과 비교하였다. 임의 두 문장, S_i와 S_j 간의 코사인 유사도 계산은 식 (5)와 같으며 유사도 값은 0과 1사이의 실수이다.

$$sim(\vec{S_i}, \vec{S_j}) = \frac{\vec{S_i} \cdot \vec{S_j}}{|\vec{S_i}| \times |\vec{S_j}|} = \frac{\sum_{k=1} s_{ki} \times s_{kj}}{\sqrt{\sum_{i=1} s_{ki}^2} \times \sqrt{\sum_{mi=1} s_{mj}^2}} \tag{5}$$

2.2 클러스터링을 이용한 요약

K-Means 클러스터링 알고리즘[9,10]은 정보검색 분야에 많이 사용된다. 이 클러스터링 방법을

사용하여 관련 문서끼리 분할하고 이중에 중요한 문장을 선택함으로써 중복성을 제거하게 된다. <그림 1>은 클러스터링 기법을 이용한 요약 알고리즘이다.

```
1. Choose k.
2. Select k initial centroids, cj, where 1≤j≤k.
3. for i = 1 to no. of sentences
4. for all j=1, 2, ···, k.
5. Compute dist(si, cj).
6. endfor
7. Select j for the minimum dist(si, cj).
8. Assign si to the jth cluster.
9. endfor
10. Recompute the new centroids for each cluster.
11. Check if (old centroids ≒ new centroids) then return.
12. else goto 3.
13. Choose the nearest sentences of the centroids in the clusters.
```

<그림 1> 클러스터링 기법을 이용한 요약 알고리즘

클러스터링 기법을 사용할 경우 문서의 내용이 많지 않을 때 문장과 문장 사이에 일치하는 용어의 수는 적다. 그렇기 때문에 문장을 이용하여 초기 센트로이드를 설정하는 것에 어려움이 있다. 본 논문에서는 2~3개의 문장을 초기 센트로이드로 설정하여 이 문제를 보완하였다. 초기 센트로이드 계산 방법은 식 (6)과 같다.

$$C_j^{initial} = \frac{\sum_i |s_i| \cdot \overrightarrow{SV_{s_i}}}{\sum_i |s_i|} \qquad (6)$$

여기서 $|s_i|$는 문장의 용어수이고 $\overrightarrow{SV_{s_j}}$는 s_i의 문장벡터이다.

2.3 완전그래프를 이용한 요약

자동요약은 사람이 한 요약과 얼마나 유사한지에 따라 정확도가 결정된다. 일반적인 인위적인 요약은 전체 내용을 파악하여 관련성이 있는 문장 중 내용을 대표하는 문장으로 요약한다.

그래프를 이용한 인위적 요약 방법으로 Salton의 Text relationship map이 있다[4]. Salton의 요약 방법은 문장 간의 관련성이 적은 문장을 제거한 후 문장 간 사슬 관계에 의해 요약을 하게 된다. 그렇기 때문에 미미한 값, 즉 관련성이 적은 문장 사이의 거리값은 적용되지 않는다. Salton의 단점을 보안한 방법으로 문장 간의 거리값을 모두 합한 '도합유사도'가 있다[12]. 그러나 이들 두 방법 모두 유사성이 높은 문장을 요약문으로 선택하는 방법을 사용하였다.

본 논문에서 제안하는 완전그래프를 이용한 요약 방법은 문서 내에 있는 모든 문장을 그래프의 꼭짓점으로 하고 각 꼭짓점 사이의 거리(문장 간의 유사도)를 계산하여 문장과 문장 사이의 선분값으로 간주한다. 문장을 선택하는 방법은 각 문장과 연결된 모든 선분값을 합하여 문장의 중요도를 계산하였다. 본 논문에서는 각 문장에 연결된 선분값의 합이 높은 순서로 문장을 선택하는 방법과 낮은 순서로 선택하는 방법을 계산하여 그 결과를 비교하였다. 식 (7), (8)은 본 논문에서 사용한 문서 요약 식이다.

$$\arg\min^{k}\left\{\sum_{j=1}^{n} sim(S_i, S_j)^2\right\} \qquad (7)$$

$$\arg\max^{k}\left\{\sum_{j=1}^{n} sim(S_i, S_j)^2\right\} \qquad (8)$$

식(7)은 각 문장에 연결된 모든 문장의 거리(유사도)를 합하여 그 값이 가장 낮은 순서로 사용자가 원하는 수, k 만큼 문장을 추출한다. 이 때 사용하는 유사도 계산은 Cosine similarity, Inner product 계산식을 사용하여 알고리즘의 성능 향상을 유도하였다.

식(8)은 식(7)과는 다르게 각 문장에 연결된 모든 문장의 거리를 합하여 그 값이 가장 높은 순서로 문장을 추출하는 방법이다. 이 경우 문장 사이의 거리 계산에 Euclidean distance를 적용하였다.

식(8)과 식(7)은 그 계산식에서 유사도를 측정하는 방법과 문서를 추출하는 방법은 다르나 두 식 모두 문장 간의 유사성이 가장 적은 문장을 추출하게 된다.

두 방법에 대한 결과는 4장에서 자세히 설명하겠다.

완전그래프기법에서 사용된 유사도 계산식은 식 (10)-(12)과 같다. 식 (10)은 Cosine similarity, 식 (11)은 Euclidean distance, 식 (12)는 Inner

product를 나타낸다.

$$sim(\vec{S_i}, \vec{S_j}) = \frac{\vec{S_i} \cdot \vec{S_j}}{|\vec{S_i}| \times |\vec{S_j}|} = \frac{\sum_{k=1} s_{ki} \times s_{kj}}{\sqrt{\sum_{l=1} s_{li}^2} \times \sqrt{\sum_{m=1} s_{mj}^2}} \quad (10)$$

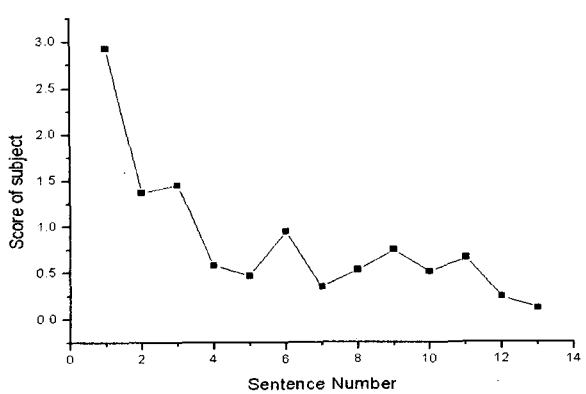
$$sim(\vec{S_i}, \vec{S_j}) = \sqrt{(\vec{S_i} - \vec{S_j})^2} = \sum_{k=1} \sqrt{(s_{ki} - s_{kj})^2} \quad (11)$$

$$sim(\vec{S_i}, \vec{S_j}) = \vec{S_i} \cdot \vec{S_j} = \sum_{k=1} s_{ki} \times s_{kj} \quad (12)$$

3. 인위적 요약

실험문서로는 20개의 Korea Times 기사를 이용하였다. 각 문서는 평균 11.5개의 문장으로 실험 문서의 형평성을 고려하여 문서 내의 문장수가 11~13개인 문서를 선택하였다.

인위적 요약에 참가한 사람은 전북대학교 인문대 대학원생 5명과 본 요약시스템 개발자를 포함한 전북대학교 공과대 대학원생 4명이다. 각 실험 대상자는 기사를 먼저 읽고 각각의 기사에 대해 가장 중요한 문장을 1개 선택하고 그 다음 문장의 중요성에 따라 차기 문장을 선택하도록 했다. 이렇게 해서 선택된 5개의 문장을 중요도에 따라 1에서 5까지 각각 점수를 부여했다. <그림 2>는 인위적 요약방법에 따라 선택된 문장들의 평균 점수변화를 보인다. <그림 2>에서는 대부분 뉴스 기사의 중요한 문장이 앞에 위치한다는 것을 보여준다.



<그림 2> 인위적 요약방법에 의해 선택된 문장들의 점수변화

4. 실험결과

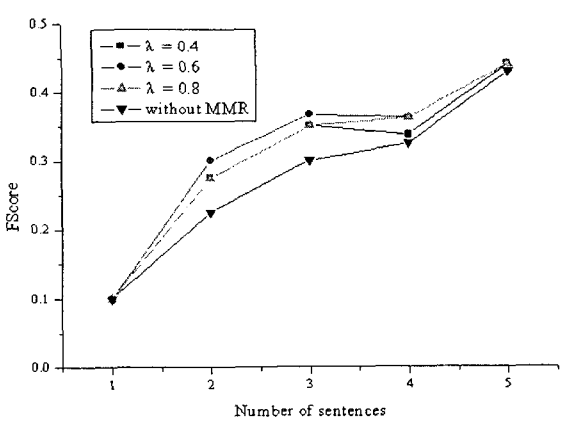
본 실험은 인위적 문서요약을 기준으로 하여 MMR, 클러스터링, 완전연결기법을 이용한 요약을 비교하였다. 비교 값으로는 재현율과 정확율의 정보를 포함하고 있는 FScore를 사용하였다[2].

$$recall = \frac{|B|}{|A+B|} \qquad precision = \frac{|B|}{|B+C|},$$
$$FScore = \frac{2 * recall * precision}{recall + precision} = \frac{|CS \cap HS|}{No.\ of\ summaried\ sentences} \quad (9)$$

여기서 CS는 자동요약으로 생성한 문장집합이고 HS는 인위적 요약의 문장 집합이다.

본 실험에서는 비교를 용이하게 하기 위하여 인위적 요약문서의 문장수와 통계 요약문서의 문장수를 같게 하였다. 따라서 재현율의 값은 정확율과 같게 되며, 아울러 FScore의 값도 재현율이나 정확율의 값과 같게 된다.

<그림 3>은 MMR기법의 임계값(λ)에 따른 요약의 성능을 나타낸다. MMR기법은 통계값만을 이용한 요약(without MMR)보다 항상 좋은 성능을 보인다. <그림 3>에서는 임계값 λ=0.6일 때 가장 좋은 요약 결과를 보인다.



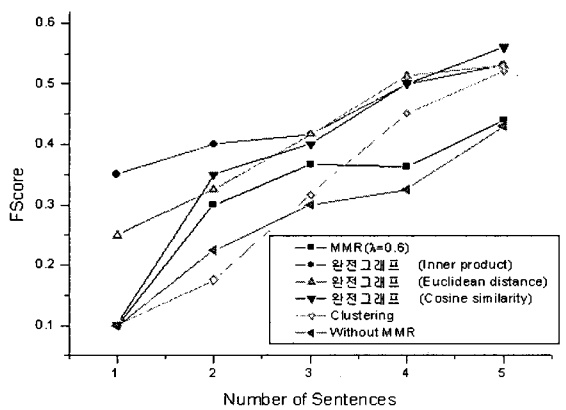
<그림 3> 임계값(λ)과 요약문장 수에 따른 FScore

<그림 4>는 요약 문장수와 요약방법에 따른 FScore값을 보여준다. 완전그래프기법은 단순 통계값을 기초로 한 'without MMR', 'MMR(λ=0.6)' 그리고 'Clusteing' 기법보다 매우 나은 성능을 보인다. 특히 Inner Product와 Euclidean Distance 유사도 계산을 이용한 완전그래프기법의 경우 1~3문장 요약에서 다른 요약방법보다 FScore값

이 매우 높게 나왔다.

검색엔진에서 요약문은 보통 2~3개의 문장으로 이루어져 있다. 원문의 길이에 비하면 매우 적은 내용이다. 일반적인 단순 통계기반 방법들은 원문의 내용보다 적은 요약일 경우 성능 저하되는 것을 볼 수 있다.

따라서 완전 그래프기법은 요약할 내용이 짧고 간결한 경우에 유리할 것이다.



<그림 4> 요약 문장수와 요약방법에 따른 FScore

Salton의 요약방법에서는 문장 간의 유사도가 높은 것을 이용하여 요약을 한다. 그러나 본 논문의 완전그래프기법에서는 문장 간의 유사도가 낮은 값을 추출하도록 설계되었다. 따라서 기존의 유사도가 높은 개념이 아닌 다른 개념으로 요약문 생성에 접근하였다.

5. 결 론

본 논문에서는 완전그래프를 이용한 문서요약 기법을 제안하였다. 이 방법은 단순한 문장 사이의 통계치를 이용한 기존의 방법보다는 문서내의 한 문장과 그 문장을 제외한 모든 문장과의 관계를 고려하였다. 또한 이 방법은 문서의 내용을 많이 포함하면서 다른 문장과는 유사성이 없는 문장을 선택하도록 함으로써 요약 성능을 높인다.

이 기법의 성능 평가를 위하여 MMR, 클러스터링 기법 등의 문서 요약과 비교하였다. 그 결과 문장 간의 의미, 즉 문맥을 적용한 완전그래프기법의 요약 성능이 가장 좋았다. 클러스터링 요약의 경우 요약 내의 문장수가 많을 때 비교적 좋은 결과를 보였다. 그러나 요약 문장수가 적을 때는 그 결과가 좋지 않았다. 그 이유는 클러스터링 기법이나 MMR 기법이 대표문장을 선택할 때 단순히 한 문장에 대한 통계적인 계산의 가중치를 기초로 문장을 선택하기 때문이다. 본 논문에서 제안한 완전그래프는 각 문장이 자기 이외의 모든 문장과의 상관관계를 고려해야 함으로 이러한 단순 통계학적 계산의 단점을 제거할 수 있다.

완전그래프기법 중 Euclidean 거리 계산을 이용할 경우, 문장사이의 거리값이 클 때, 즉 각 문서간의 연관성이 적고 용어의 정보를 많이 포함한 내용을 요약문으로 선택한다. 따라서 문서내의 문장수가 적은 경우, 특히 문장 간의 연관관계가 적은 경우 유리하다.

반면에, 문서의 내용이 많지 않고 각 문장의 길이가 짧은 경우 완전그래프기법의 유사도 계산 중 Inner product와 Cosine similarity 계산방법에는 문제점이 있다. 왜냐하면 문서에 있는 각 문장이 가지고 있는 정보가 적기 때문 문장 사이의 유사도를 구하기가 힘들기 때문이다. 이 경우 사전이나 유사어, 관련어를 이용한 용어확장 방법이 필요하다. 이런 방법을 이용하면 좀 더 정확한 요약문을 얻을 수 있을 것이다.

완전그래프기법을 이용하면 문서의 내용을 적은 문장으로 요약할 경우 요약 성능이 다른 방법보다 매우 우수하다. 따라서 웹문서와 같은 적은 수의 요약에서는 완전그래프기법 요약처럼 문맥을 이용한 요약방법을 사용하면 보다 정확한 요약문을 생성할 수 있을 것이다.

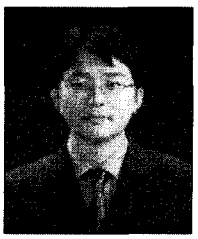
참 고 문 헌

[1] 강상배, 한국어 문서의 통계적 정보를 이용한 문서요약 시스템 구현, 부산대학교, 전자계산학과 , 석사 학위 논문, 1998. 2.

[2] Jade Goldstein, Mark Kantrowitz, Vibhu Mittal, and Jaime Carbonell, Summarizing Text Documents: Sentence Selection and Evaluation Metrics, In Proceedings of ACM-SIGIR'99, Berkeley, CA, August 1999.

[3] Inderjeet Mani and Mark Maybury. Advances in Automatic Text Summarization. MIT Press, 1999.

[4] G.Salton, A.Singhal, M.Mitra, and C.Buckley, Automatic Text Structuring and Summarization.
[5] W. Kraaij, M. Spitters, and M. van der Heijden. Combining a mixture language model and naive bayes for multi-document summarisation. In Working notes of the DUC2001.
[6] Jaime Carbonell and Jade Goldstein, "The use of MMR, diversity-based reranking for reordering documents and producing summaries," in Proceedings of the 21st ACM-SIGIR International Conference on Research and Development in Information Retrieval, Melbourne, Australia, 1998.
[7] Jade Goldstein, Mark Kantrowitz, Vibhu Mittal, and Jaime Carbonell, Summarizing Text Documents: Sentence Selection and Evaluation Metrics, In Proceedings of ACM-SIGIR'99, Berkeley, CA, August 1999.
[8] 유준현, 변동률, 박순철, "단일 문서의 인위적 요약과 MMR 통계요약의 비교 및 분석", 대한전자공학회논문지, 제41권 CI편 제2호, 2004년 3월.
[9] A. Leuski and J. Allan. Improving interactive retrieval by combining ranked lists and clustering. In Proceedings of RIAO'2000, pages 665--681, April 2000.
[10] 오형진, 고지현, 안동언, 박순철, "색인어 가중치 부여 방법에 따른 K-Means 문서 클러스터링의 LSI 분석", 정보처리학회논문지, 제10-B권 제7호, 2003년 12월.
[11] Inderjeet Mani and Eric Bloedorn, Summarizing Similarities and Differences Among Related Documents, Information Retrieval 1 (1-2), pages 35-67, June 1999.
[12] 김준홍, 도합유사도를 이용한 한국어 추출요약 시스템, 한국해양대학교, 컴퓨터공학과, 석사 학위 논문, 2000. 8.



유 준 현 (Jun Hyun Lyu)

- 2002년 2월 : 전북대학교 공과대학 (공학사)
- 2004년 2월 : 전북대학교 정보통신학과 (공학사)
- 2004년-현재 : LG Electronics Co. Researcher
- 관심분야 : 정보검색, 데이터베이스

박 순 철 (Soon Cheol Park)

- 정회원
- 1979년 2월 : 인하대학교 공과대학 (공학사)
- 1991년 12월 : (미국)루이지아나 주립대학 (전산학박사)
- 1991년-1993년 : 한국전자통신연구원 근무
- 1993년-현재 : 전북대학교 전자정보공학부 교수
- 관심분야 : 정보검색, 온톨로지