전자정부내 의미기반 기술 도입에 따른 기능 및 정책 연구

(Research on Function and Policy for e-Government System using Semantic Technology)

장 영 철*
(Young-Cheol Jang)

요 약 본 논문에서는 전자정부 시스템의 대 국민 사용성과 효율성을 증진시키기 위한 의미기반 문서분류 방법(CoWDC)을 제시한다. 기존 의미기반 문서분류 방법에서 많은 양의 키워드들의 계층적 컨셉들을 이용하는 것을 지양하고 사용자들이 사용하는 키워드들 간의 관계를 중심으로 문서를 분류한다. 즉, 문서의 컨텍스트(context)에 근거하여 깊고 정확한 의미를 키워드 간 관계를 분석하여 적은 양의 정보로 효율적인 문서분류를 하게 된다. 이를 위해 제안한 CoWDC(Concept Weight Document Classification) 시스템은 기존의 시소러스/온톨로지의 의존도를 줄이고 키워드 관계, 관계의 경중 고려, 상하위 개념으로 변환 등을 통한 실험과 평가가 이루어졌다. 전자정부 시스템의 구조 및 특징 분석을 통해 CoWDC 실험 결과는 대국민 서비스 향상을 위해 매우 필요함을 인지하고 이를 접목하기 위한 기술적, 정책적 제언을 제시하였다. CoWDC를 통해 의미기반 검색기술의 우수함을 입증하였고 이는 전자정부 시스템의 지식베이스 구축, 운영체제의 운용, 시소러스의 구성 등의 과정에서 체계적으로 통합 운영되어야 한다.

핵심주제어 : 문서 분류, 개념 가중치, 세만틱 웹, 시소러스, 전자정부

Abstract This paper aims to offer a solution based on semantic document classification to improve e-Government utilization and efficiency for people using their own information retrieval system and linguistic expression. Generally, semantic document classification method is an approach that classifies documents based on the diverse relationships between keywords in a document without fully describing hierarchial concepts between keywords. Our approach considers the deep meanings within the context of the document and radically enhances the information retrieval performance. Concept Weight Document Classification(CoWDC) method, which goes beyond using existing keyword and simple thesaurus/ontology methods by fully considering the concept hierarchy of various concepts is proposed, experimented, and evaluated. With the recognition that in order to verify the superiority of the semantic retrieval technology through test results of the CoWDC and efficiently integrate it into the e-Government, creation of a thesaurus, management of the operating system, expansion of the knowledge base and improvements in search service and accuracy at the national level were needed.

Key Words : Document Clustering, Concept Weight, Semantic Web, Thesaurus, e-Government

1. 서 론

세계의 선진 전자정부는 정보화 사회의 주축 및 파트너 축으로서 국민(수요자) 위주의 서비스 질을 향상하고 경제, 사회활동에 전자정부의 정보가 정보경제를 창출하는 핵심 기반요소가 되고 있다. 우리나라도 1987년부터 주민, 부동산, 자동차 등

* 경민대학 멀티미디어방송과

국가기관 전산망 사업을 시작으로 2001년 전자정부 11대 과제를 선정하여 민원서비스 혁신을 위한 G4C(Government for Citizen) 구축, 2003년 전자정부 로드맵 작성, 2008년까지 민원업무 온라인화 85%라는 목표로 정진하고 있다. 이러한 결과로 일부 기능별·부처별 단위업무 정보화는 고도화 단계에 진입하였지만, 국민의 전자정부 활용률은 35.1%를 보이고 있다. 국민들의 눈높이에 맞는 원스톱 서비스, 검색 성능 향상, 사용자중심의 정보재구성 및 의미기반 정보 서비스 체계를 구축하여 전자정부의 활용성 및 효율성을 높여야 할 것이다[15,16].

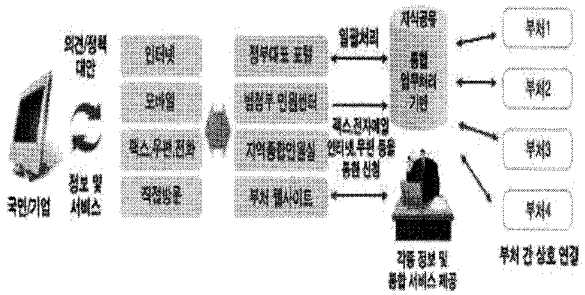
사람이 갖고 있는 지식체계를 모방하며 인간과 유사하게 문서를 의미기반으로 분류하는 접근은 복잡하고 모호한 전자정부 검색 환경 하에서 검색의 성능을 혁신적으로 향상시키게 되며 사용자 이용 프로세스에 친화적인 검색형태를 제공하게 되어 대국민 접근성 및 활용성을 향상시키는 중요 기반이 된다. 전자정부 시스템에 이러한 시맨틱 기술을 사용한 정보 문서분류 및 검색 체계를 구축하는 일은 기술적인 문제뿐만 아니라 이를 지원하는 법제도 및 정책적인 지원시스템이 균형적으로 갖추어져야 한다.

본 논문에서는 전자정부에 의미기반 기술을 적용한 지식정보 체계 및 문서분류 체계를 구축하면서 이러한 요소들의 관계와 중요성을 기술하고 이를 구현하기 위한 전략 및 정책적 개선방안을 제시한다.

2. 전자정부와 의미기반 문서분류

2.1 국내외 전자정부의 특성

대한민국 전자정부(www.korea.go.kr), 국가지식포털(www.knowledge.go.kr), 통합전자민원창구(www.egov.go.kr) 등을 통해서 본 국내 전자정부는 문서유통체계 고도화 및 중앙행정기관 자료관 시스템 구축, 국회의정 참여 포털 구축, G4C 및 공동이용 확대, 온라인 국민 참여 포털 시스템 구축 등에 중점을 두고 사업을 진행해오고 있다[11].



(그림 1) 국내 전자정부 구현 모델

하지만 문서 분류체계의 수가 단순하고 사용자 업무 중심으로 설계되어 있지 않아 정보검색이 불편하고 요약설명 기능이 부족하여 포털로서 관문 기능이 부족하다. 또 성급한 표면상의 통합으로 인해 메타데이터의 보강 및 관계 연결이 조정되지 않고 불용어 처리가 의미기반으로 되지 않아 미관련 문서가 너무 많이 검색되고, 띄어쓰기에 따라 다른 검색결과가 나오고, 결과내 검색, 질의자의 의도 및 목적 표현의 중요도가 반영되지 않아 사이트의 이용률이 떨어지고 있는 실정이다[15,16].

미국의 전자정부(www.firstgov.gov)는 시민 중심, 결과 중심, 시장 기반의 3가지 원칙하에서 온라인 검색시스템인 DgPort를 통해 전자정부 용어에 대한 시소러스와 검색 결과를 시각화 하고 주제 별 분류, 민원 서비스 소재 정보 바로가기 등의 특징이 있다[12].

영국의 전자정부(www.direct.gov.uk)는 전자상거래전략, 인터넷확산전략, 전자정부전략을 병행하여 추진하며(UK-Online) 국민의 일상생활 중 주소변경일괄통지서비스, 국내외 여행, 출산, 자동차 관련정보 등이 제공되고 있으며 이용자들의 지속적인 수요조사 결과를 반영하여 새로운 서비스를 제공하는 피드백 메커니즘과 AskJeeves 방식의 관련 정보 추천 기능, 사용자와 시스템 간의 어휘갭(Lexical Gap)을 줄인 관련 정보 제공 기능이 있다. 하지만 아직 관련 정보를 민원 프로세스 내에서의 연계성까지는 고려하지 못하고 있다[11].

캐나다 전자정부포털(www.canada.gc.ca)은 국

민을 위한다는 정책 철학위에서 개인 맞춤형 기능과 정부기관 웹사이트를 통합 검색하여 결과를 제공하고 'Search'기능을 최상위 매뉴로 배치하여 검색중심의 서비스를 제공하고 있다. 또한 제목 일치성과 사용자가 많이 이용한 사이트를 고려한 검색결과를 제공하고 있다. 또 프라이버시 보호원칙을 적극 반영하고 사용자가 편리한 방식(이메일, 팩스, 전화, 대면, 온라인 등)으로 서비스를 제공하고 있다[13].

2.2 의미기반 문서분류

국내 포털사이트 네이트닷컴(Nate.com)은 유의어 사전과 하위어 사전을 기본 지식베이스로 활용하고, 조달청의 상품검색에서는 상품들의 정보를 온톨로지화하여 의미기반 검색을 수행하는 등 검색 성능 및 질의 향상을 위해 의미기반 검색을 시도 하고 있고 정보가 폭증함에 따라 문서분류의 자동화, 표준화, 지능화의 필요성은 더욱 절실해지고 있다[8].

본 논문에서 문서분류에 사용하는 의미기반 접근은 기존의 시맨틱웹 기술로 표현된 의미표현의 깊이를 더 확장하고 문서내 다양한 키워드 관계하에서도 주제어 선정이 정확히 이루어지게 하기 위하여 객체의 상속관계 운용기술과 인공지능 개념학습 분야의 일반화, 세분화 이론을 응용한 문서분류 접근을 말한다. 전자정부 내의 검색시스템의 의미화와 고도화를 위해, 문서 내 핵심이 되는 키워드들을 충분히 파악하고, 이들 키워드들을 중심으로 관련 개념들을 선정한 후 제안한 프로세스에 따라 개념들의 의미 범주(concept space)를 조정함으로써(상속-역상속) 의미기반으로 문서의 주제 선정 및 문서분류를 수행하는 방법을 제안한다[1,2].

3. 의미기반문서분류(CoWDC)

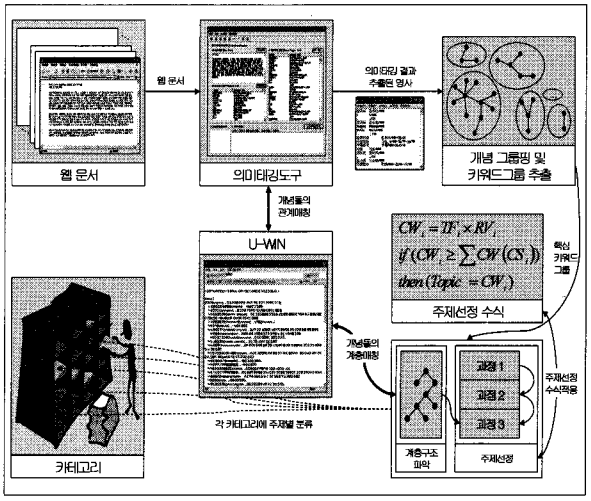
국민의 사용 절차와 정부 행정업무의 전문성에 부합하는 전자정부 체계를 구축하기 위해서는 문서를 분류하는 체계가 기존의 키워드와 단순한 시소러스/온톨로지 사용을 넘어서 다양한 개념들의 계층관계 까지 충분히 고려된 분류가 되어야 함을 주장하고 이를 위해 CoWDC(Concept Weight Document Classification) 방법을 제안하고 평가하였다.

CoWDC는 1) 문서내 키워드들의 빈도수 2) 시소러스 체계에서 키워드들 간 관계의 중첩빈도 3) 개념 계층(concept hierarchy)내에서 정·역 방향의 개념가중치 상속을 기반개념으로 사용한다[3,7].

3.1 키워드 그룹 및 주제선정

개념가중치적용 의미기반 문서분류 방법 CoWDC 는 크게 2개의 과정으로 구성된다. 첫째는 문서에 포함된 명사개념들의 출현횟수(TF : Term Frequency)와 관계횟수(RV : Relation Value)를 이용한 문서의 핵심이 되는 키워드 그룹을 추출해 내는 과정이고, 둘째는 핵심 키워드 그룹내에서 키워드 빈도수(TF)와 관계수(RV)를 사용한 개념가중치(CW: $CW= TF\times RV$, Concept Weight)를 이용하여 주제(top-keyword)를 선정하는 과정이다.

[그림 2]는 CoWDC를 통해 문서를 분류하는 전 과정을 보이고 있다.



(그림 2) CoWDC의 의미기반 문서분류 전 과정

"문장에서 주제에 가까운 개념일수록 출현 및 주변개념과의 관계 횟수가 많다"는 정리에 근거하여 TF와 RV요소를 모두 고려한 CW를 이용하여 주제를 선정하는 방식이다.

<표 1> 문서의 주제 선정 과정

경우	조 건	예 외
case 1	특정 개념의 CW값이 다른 개념들의 CW값의 합보다 클 때, 특정 개념을 문서의 주제로 판단	만족하지 않을 경우 case 2 진행
case 2	하위에 있는 개념으로 상위개념의 CW값을 상속하여 case 1 적용	만족하지 않을 경우 case 3 진행
case 3	상위에 있는 개념에 하위 개념의 CW값들을 역상속하여 case 1 적용	case 3의 조건은 default 경우임

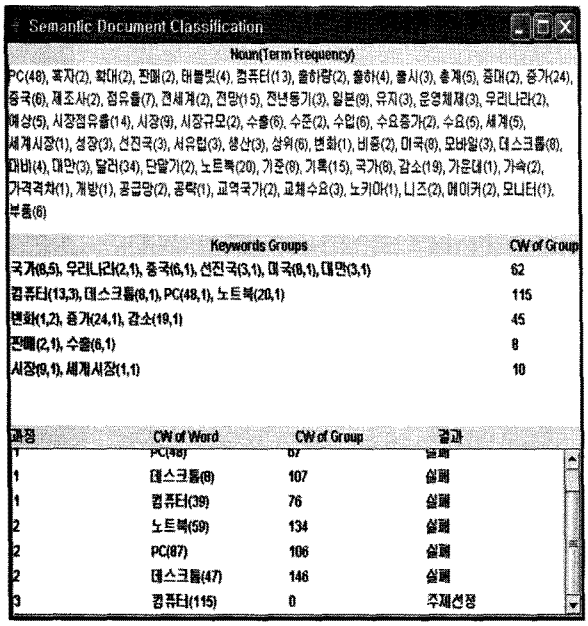
3.2 CoWDC 의 효율성 분석

개념가중치적용 의미기반 문서분류 방법 CoWDC의 효율성을 입증하기 위해, 전자정부의 국가지식포털에서 검색된 문서를 이용한 실험을 기술하고, 온톨로지 기반 분류방법 및 베이지안 분류방법과 비교하였다.

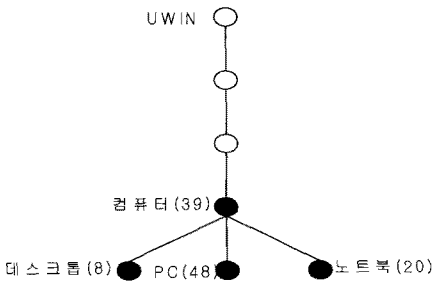
[그림 3]에서 case 1(어느 한 개념의 CW가 월등히 높을 경우), case 2(상위개념을 상속받는 경우)는 실패하고 case 3(상위개념이 하위개념을 모두 포함하는 경우)에 해당한다.

<표 2> 키워드 그룹 선정

그룹	키워드(TF, RV)	cwG_j
1	국가(8,5), 우리나라(2,1), 중국(6,1), 선진국(3,1), 미국(8,1), 대만(3,1)	8 · 5+2 · 1+6 · 1+3 · 1 +8 · 1+3 · 1 = 62
2	컴퓨터(13,3), 데스크톱(8,1), PC(48,1), 노트북(20,1)	13 · 3+8 · 1+48 · 1+20 · 1 = 115
3	변화(1,2), 증가(24,1), 감소(19,1)	1 · 2+24 · 1+19 · 1 = 45
4	판매(2,1), 수출(6,1)	2 · 1+6 · 1 = 8
5	시장(9,1), 세계시장(1,1)	9 · 1+1 · 1 = 10

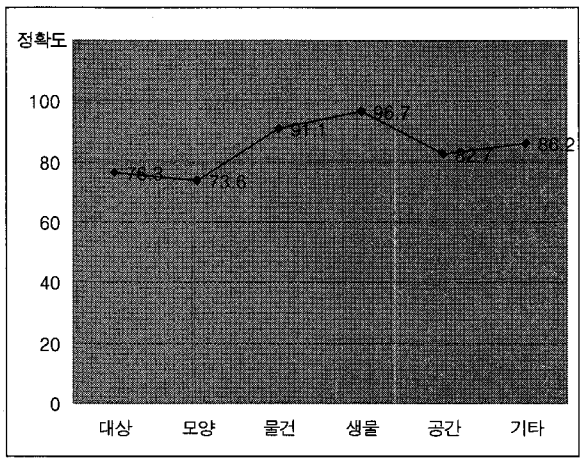


(그림 3) 주제선정 실험결과



(그림 4) 선택된 키워드 그룹과 시소러스(U-WIN)와 매핑한 계층구조

CoWDC를 통한 실험에서 추출된 문서분류와 실제 사람이 생각하는 문서분류를 비교하여 일치한 결과에 대한 통계를 낸 결과 84.4%의 정확도를 보였다.



(그림 5) 계층영역별 실험결과

하지만 '대상', '모양', '공간' 등의 도메인은 계층을 구성하는 개념들이 자주 사용되지 않고 일부 계층구조를 이루지 않는 개념(상/하위 계층을 이루지 않고 독립적으로 존재)들이 존재하여 [그림 5]와 같이 도메인에 따라서 정확도의 차이가 컸다. 이는 시소로스 각 도메인 내에 개념들이 얼마나 체계적으로 작성되느냐가 문서분류에 영향을 미친다는 결론을 도출할 수 있다.

CoWDC 의 효율성 비교를 위하여 Sinka[4]의 연구에서 제공되는 14가지 분류범주하에서 베이지안 방식과 온톨로지 방식으로 실험하였다. 각 분류범주 별로 100개의 문서학습을 통하여 각 분류범주에 관련된 개념들을 학습한 후 3600개의 실험문서 표본을 이용하여 분류를 시도하였다. CoWDC를 통한 문서분류가 온톨로지 기반 분류방법(지식베이스유형)보다 0.9%, 베이지안 분류 방법(벡터유형)보다 11.4% 향상된 정확도를 보였다.

4. CoWDC 기반의 전자정부 시스템

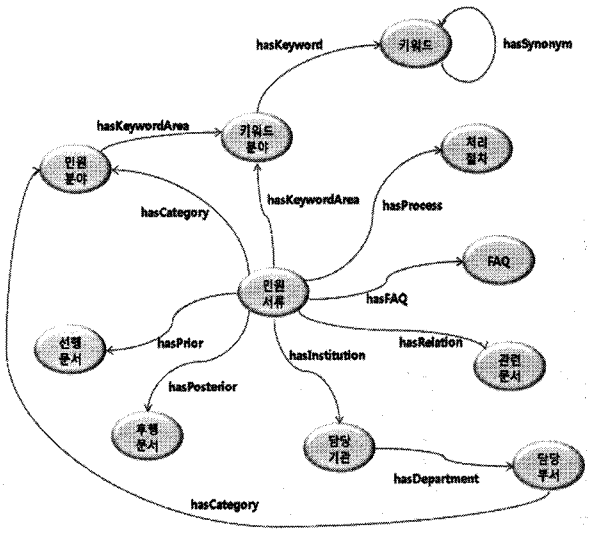
의미기반 접근이 기존의 벡터 나 지식베이스 접근 보다 우월함을 근거로 전자정부에 이를 구현하기 위해서는 다음의 기술적 확장과 정책적 고려가 필요하다.

4.1 기술적 제언

문서분류의 정확도에 시소러스의 영향이 중요하므로 실제 문서에서 사용되는 일반 개념, 외래어 및 전문 용어의 개념, 정의된 개념들의 관계 등을 모두 파악할 수 있도록 개념 확장이 우선적으로 필요하다. [그림 6]은 지식베이스 확장을 위해 "민원서류" 중심의 시소러스 설계 사례를 보여주고 있다. 또 의미기반 문서분류 방법을 실제 대한민국 전자정부의 검색서비스에 적용하여 진정한 국민을 위한 G4C로 거듭나기 위해서는 검색 차원에서 복합명사 분할, 검색어의 중요도에 의한 랭킹, 주제별 검색결과의 동적랭킹, 검색 결과의 자동 분류 등의 개선사항들이 요구된다.

4.2 정책적 제언

전자정부에 의미기반 문서분류를 도입하기 위해서는 국가적 차원의 시소러스 구축 전담기구가 가이드라인을 가지고, 시소러스간 통합, 전문용어 표준화, 국내외 연구기관과 연동을 지속적으로 연구지원해야 하고, 전자정부 전용 검색엔진 개발, 검색 서비스 평가를 통한 개선, 검색 서비스 가이드라인이 제시되어야 한다. 또 민원 프로세스가 구축



(그림 6) "민원서류"를 중심의 시소러스 설계 사례

되고, 의미기반 기술에 바탕을 둔 전자정부포털이 구성되고 이 기술의 상용화를 정부가 민간과 파트너쉽으로 적극 지원해야 한다. 전자정부내 지식정보가 국민의 삶의 질 향상(경제, 건강, 복지 등)을 위해 재사용 되도록 구성되어야 국민들의 이용율이 향상될 수 있을 것이다.

5. 결 론

본 논문에서 제안된 개념가중치적용 의미기반 문서분류 CoWDC는 기존의 베이지안 접근방법의 많은 문서 학습의 어려움(학습시간, 기준문서준비), 비학습 영역의 현저한 성능저하, 구체적인 관계 기술의 어려움을 한국어 시소러스와 연계한 프로세스를 설계함으로써 극복하였고 지식베이스 기반 접근방법이 모든 관계 표현을 완벽히 구성하기 어렵고 사전 노동력이 많이 든다는 단점을 객체의 상속 관련 특성 및 개념가중치(CW)를 사용하여 개선함으로써 문서분류에 의미부여 특징을 강화하게 되었다. 또 CoWDC 구성 및 실험과정에서 의미기반 문서분류 특성을 강화하는데 지식베이스의 체계와 이의 지속적인 운영 관리 체계가 중요한 요소임을 인식하고 표준화되고 통합된 전문용어, 시소러스 체계의 구축 및 관리 활용을 국가차원에서 지원해야 한다는 정책들을 제안하였다. 결국 전자정부의 고도화와 대국민 활용성 향상을 위해서는 삶의 질 향상에 필요로 하는 내용이 의미기반 방식으로 서비스되고 이를 지속적으로 제공하고 기술 환경의 변화에 따라 진화하며 국민 친화적이 되어야 함을 본 논문의 최종 결론으로 맺는다.

참 고 문 헌

[1] C. Buckley, G. Salton and J. Allan(1994). "The Effect of Adding Relevance Information in a Relevance Feedback Environment", In Proc. 17th ACM SIGIR International Conference on Research and Development in Information Retrieval, pp. 292-298.

[2] Hyunjang Kong, Myunggwon Hwang, Gwangsu Hwang, Jaehong Shim, Pankoo Kim(2006). "Topic Selection of Web Document Using Specific Domain Ontology", MICAI2006, LNAI4293, pp.1047-1056.

[3] J.S. Shin, J.H. Kwak and C.H. Lee(1999). "Automatic Classification of Web Documents with Word Accordance of Degree using Probability Model", Proceedings of ICOIN 13, Jan., pp. 6A-3.1-6A-3.4.

[4] M.P.Sinka and D.W.Corne(2002). "A large benchmark dataset for web document clustering," Soft Computing Systems: Design, Management and Applications, Frontiers in Artificial Intelligence and Applications, Vol.87, pp.881-890.

[5] Marko Balabanovic and Yoav Shoham(1995). "Learning Information Retrieval Agents: Experiments with Automated Web Browsing", Proc. 1995 AAAI Spring Symposium on Information Gathering from Heterogeneous, Distributed Environments, Stanford, AAAI press, March

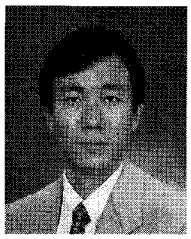
[6] R.Hanson, J.Stutz and P.Cheeseman(1991). "Bayesian Classification Theory", Techinical Report FIA-90-12-7-01, NASA Ames research Center, AI Branch.

[7] Salton, G., and Buckey(1987). "Term weighting approaches in automatic text retrieval", Technical Report 87-881, Cornell University, Department of Computer Science.

[8] T. Berners-Lee, J. Hendler, and O. Lassila. (2001). "The Semantic Web". Scientific American, May 2001.

[9] 고수정, 이정현(2001). "Apriori-Genetic 알고리

즘을 이용한 베이지안 자동 문서 분류", *정보처리학회 논문지 B*, Vol.01, No.01, p.1~12.
[10] 고영만(2006). "시소러스 기반 온톨로지에 관한 연구", 성균관대학교 정보관리연구소, 정보관리 제5집, pp.5~22.
[11] 한국전산원(2006). "국내외 정보화동향: 전자정부 동향". NCAWeekly No.22.
[12] 한국전산원(2005). "미국 전자정부의 현황 I", 전자정부 동향 2005-04.
[13] 한국전산원(2005). "캐나다 전자정부의 현황 I", 전자정부 동향 2005-03.
[14] 한국전산원(2005). "호주 전자정부의 현황 I", 전자정부 동향 2005-03.
[15] 한국정보사회진흥원(2006). "2006 국가정보화 백서 제1편 지식정보사회 실현".
[16] 행정자치부(2006). "차세대 전자정부 10대 어젠다".
[17] 허준희, 최준혁, 이정현, 김중배, 임기옥(2001) "문서의 주제어별 가중치 부여와 단어 군집을 이용한 한국어 문서 자동 분류 시스템", *정보처리학회논문지 B*, 제5호, pp.447-454.



장 영 철 (Young-Cheol Jang)

- 정회원
- 1987 한양대학교 수학과(이학사)
- 1995 건국대학교 컴퓨터공학과 (공학석사)
- 1998 건국대학교 컴퓨터학과 (공학박사)
- 1996~현재 경민대학 멀티미디어방송과 교수
- 관심분야: 디지털경제정책, 첨단영상, 인공지능