특허 문서 텍스트로부터의 기술 트렌드 탐지를 위한 언어 모델 및 단서 기반 기계학습 방법

(A Language Model and Clue based Machine Learning Method for Discovering Technology Trends from Patent Text)

전 영 실 † 김 영 호 † 정 윤 재 †† 류 지 희 †† 맹 성 현 †††
(Yingshi Tian) (Youngho Kim) (Yoonjae Jeong) (Jihee Ryu) (Sung-Hyon Myaeng)

요 약 특허 문서는 과학기술 발전을 탐지하고 기존 트렌드를 이해함으로써 미래의 트렌드를 예측하는데 유용한 자원이다. 본 연구에서는 단위 기술을 "문제점"과 "해결방법"으로 구성되어 있다고 보고, 언어적 단서(linguistic clue)와 언어 모델(language model)을 결합한 혼합 모델을 사용하여 이들에 해당하는 의미 핵심문구(semantic keyphrase)를 찾고, 의미 핵심문구로 표현되는 단위 기술을 추출하였다. 추출된 결과에 근거하여 비지도 학습(unsupervised learning) 방법으로 과학기술들의 트렌드를 발견하는 새로운 접근방법(Technological Trend Discovery, TTD)을 제안한다. 실험 결과에 따르면 본 연구에서 제안한 방법으로 과학 기술을 나타내는 의미적 핵심 문구를 추출하는데 77%의 R-정확률을 달성하였고 결과적으로 의미있는 과학기술 트렌드를 발견할 수 있었다.

키워드 : 특허, 텍스트 마이닝, 과학기술 트렌드 탐지, 의미 핵심문구 추출

Abstract Patent text is a rich source for discovering technological trends. In order to automate such a discovery process, we attempt to identify phrases corresponding to the problem and its solution method which together form a technology. Problem and solution phrases are identified by a SVM classifier using features based on a combination of a language modeling approach and linguistic clues. Based on the occurrence statistics of the phrases, we identify the time span of each problem and solution and finally generate a trend. Based on our experiment, we show that the proposed semantic phrase identification method is promising with its accuracy being 77% in R-precision. We also show that the unsupervised method for discovering technological trends is meaningful.

Key words : Patent, textual-data mining, technological trend discovery, semantic keyphrase extraction

· 본 연구는 지식경제부 및 정보통신연구진흥원의 IT핵심기술개발사업의 일환으로 수행하였음[2008-F-047-01, Urban Computing Middleware 기술 개발]

† 비 회 원 : 한국과학기술원 정보통신공학과
yingshi424@kaist.ac.kr
yhkim@kaist.ac.kr
†† 비 회 원 : 한국과학기술원 전산학과
hybris@kaist.ac.kr
zzihee5@kaist.ac.kr
††† 종신회원 : 한국과학기술원 정보통신공학과 교수
myaeng@kaist.ac.kr

논문접수 : 2008년 10월 22일
심사완료 : 2009년 3월 9일

정보과학회논문지: 소프트웨어 및 응용 제36권 제5호(2009.5)

1. 서 론

일반적으로, 과학기술에 관한 문서는 분야별로 봤을 때 시간 흐름에 따라 세부적인 주제나 트렌드가 바뀌는 것을 알 수 있다[1]. 예를 들면 과학논문 말뭉치에서는 논문 발행년도를 기준으로 과학기술의 트렌드의 예측이 가능하고, 사건 탐지 및 추적(Topic Detection and Tracking)[2] 연구에서는 사건이 발생한 시점으로부터 시간 흐름에 따른 사건의 변화를 추적할 수 있다. 그러므로 특정 도메인 텍스트 데이터에서 주요 주제를 추출함으로써 의미 있는 과거의 기술 트렌드를 추출하고, 미래의 트렌드를 예측하는 것이 가능하다. 텍스트 마이닝 기법을 사용하여 트렌드 탐지와 관련된 많은 연구가 기존에 진행되어 왔으나 대부분이 뉴스 문서에서의 사건 탐지에 관한 연구에만 국한되어 있다[1,2].

최근에는 방대한 특허 문서를 가지고 과학기술의 변화를 탐지하는 연구들이 많아지고 있다[3-5]. 기존의 과학기술 탐지와 관련된 연구들에서는 데이터 마이닝 기법을 적용하여 특허망(patent network)을 구축함으로써 새로운 기술을 탐지하는 도구로 삼았다[6,9]. 그러나 기존 방법들은 다음과 같은 문제점들을 가지고 있다. 첫째, 특허망을 이용한 방법은 특허들의 연관관계를 이용하였는데 단순한 연관정보만으로는 사용자에게 유용한 정보를 제공하기가 어렵다. 또한 특허 검색 결과 너무 많아서 정보 과잉 문제가 생긴다. 이 문제를 해결하기 위해 [7,8]에서는 특허망을 기반으로 한 특허 분석 방법을 이용하였으나 특허 문서들에서 추출된 핵심 개념들 간의 연관성이 모호하거나 구축된 특허망이 복잡하여서 일반 사용자의 경우 중요한 개념을 탐지해내기가 어렵다[7]. [8]에서 제시한 예를 보면 "universal PnP GPS"이라는 개념은 "원격 조정 시스템(remote control system)"이라는 개념과 연관되어 있다고 말할 수 있으나 두 개념이 어떠한 구체적인 관계를 갖는지에 대해서는 알 수 없다. 만약 "universal PnP GPS"이라는 개념이 "원격조정시스템(remote control system)"이라는 개념을 해결하는 방법이라는 의미를 부여한다면 관계가 명확해져서 사용자들에게 보다 유용한 정보가 될 것이다.

둘째, 특허망을 구축하는 작업은 특정 도메인에 대한 많은 지식을 필요로 한다. 기존 연구들을 살펴보면 [8]에서는 전문가의 도움을 받아 특허망을 구축하였고, PATExpert[5]는 의미 연결을 위한 온톨로지를 사용하였다. 또한 [4]에서 제시한 기술 지도(technology map)도 전문가의 도움을 받아 특허망을 구축하였다. 이러한 방법론은 특정 도메인 지식을 필요로 하기에 다른 여러 도메인에 적용하기 어려우며 통계적 방법을 통하여 일반화 시키기도 어렵다. 이를 해결하기 위하여 [9]는 통계 학습 방법 대신 단어 빈도수를 이용하여 회로 장치(circuit device)와 관련된 문서들을 분석하였고 [6]에서는 유한 상태 기계(finite-state machine) 기반의 핵심 개념 추출 방법을 제안하였다. 상세한 내용은 5장에서 더 설명하기로 한다.

기존 방법들의 단점을 보완하기 위해서 본 연구에서는 특허 문서가 다루고 있는 기술적 "문제점"과 "해결 방법"에 초점을 맞추어 이를 지칭하는 의미 핵심문구들을 자동으로 추출함으로써 언급된 과학기술 내용을 인식하고, 이를 연결하여 과학기술 트렌드를 탐지하는 방법(Technological Trend Discovery, TTD)을 제안한다.

본 연구에서는 단위 과학기술을 특정 기간 특정 도메인에서의 문제점(예: recognizing spoken language)과 해결 방법(예: 언어 모델, language model)의 조합으로 정의한다.

예를 들어 음성 인식 도메인에서의 과학기술 트렌드를 탐지하기 위해 수천 개의 문서들로부터 중요한 과학기술의 내용을 탐지하고 특정 기간 내의 주요 기술 트렌드를 탐지하는 것은 쉬운 일이 아니다. 하지만 본 연구에서 제안한 방법을 이용하여 그림 1과 같이 특허 문서들을 요약하여 표현한다면 사용자들이 대량의 기술 문서로부터 기술 동향 정보를 효율적으로 얻을 수 있게 된다.

그림 1을 보면 사용자는 음성 인식 기술이 (1) 1980년대의 동적 프로그래밍(Dynamic programming) 방법을 이용한 기술에서 1990년대에는 은닉 마르코프 모델(Hidden Markov Model)방법을 이용한 기술로 바뀌었고 (2) 1990년대 초에 화자 검증(Speaker verification)에 관한 문제점들이 제기되기 시작하였고 Dynamic Time Warping 방법이 도입되었으며, (3) 2000년대에는 언어 모델(language model) 방법론이 음성 인식과 화자 검증에 관한 문제를 해결하는데 쓰였음을 한 눈에 파악할 수 있을 것이다.

본 연구에서 제안하는 TTD 방법은 (1) 과학기술 내용 탐지를 위한 의미 핵심문구 추출(태스크 1)과 (2) 과학기술 트렌드 탐지(태스크 2)로 구성된다. 태스크 1에서는 언어적 단서와 확률모델을 사용하였다. [2]에서 제시한 확률모델은 특정 기간 내의 핵심주제를 탐지하는 방법으로 본 연구와 흡사하지만 한 단어 모델(unigram model)만을 사용하였다. 그러나 본 연구에서는 복수 단어(multi-word)까지 고려한다.

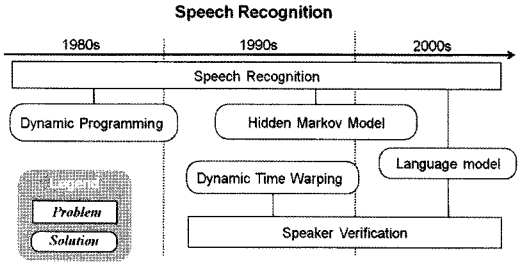

그림 1 음성 인식 기술 트렌드

또한 본 연구에서 정의한 문제점 및 해결 방법 범주는 [2]에서의 범주와 성격이 엄연히 다르다. 본 연구에서는 언어적 단서를 이용한 확률적 프레임워크를 만들었으며 패턴(즉, 언어적 단서) 가중치를 통계적 학습기의 자질로 사용함으로써 언어적 단서의 적응성(adaptability)을 향상시켰다. 태스크 2에서는 특정 기간 동안의 핵심과학기술을 찾고 과학기술 간의 연관관계를 정의하여(2장) 과거의 과학기술 트렌드를 탐지할 수 있도록 하였다.

본 연구에서는 1976~2003년까지 출원된 음성 인식 관련 미국 특허 데이터를 사용하여 제안한 시스템에 대해 평가를 진행하였다. 실험 결과를 통해서 본 연구에서 제안한 방법론이 기존 음성 인식 분야의 과학기술 정보를 정확하게 추출하고 특정 기간 동안의 의미 있는 과학기술 트렌드를 탐지할 수 있음을 알 수 있다.

본 논문의 구성은 다음과 같다. 2장에서는 과학기술 트렌드 탐지를 위한 태스크를 공식화 하고, 3장에서는 제안한 방법론 및 시스템에 대해 설명하며, 4장에서는 시스템의 실험 결과에 대해 다양한 측면에서 평가하고 분석한다. 5장에서는 기존의 관련 연구에 대해 설명하고 마지막으로 6장에서는 제안한 방법의 특징과 장단점을 정리한 후 향후 연구 방향을 제안한다.

2. 태스크 정의

이 장에서는 Technological Trend Discovery(TTD) 태스크에 관련된 새로운 개념들을 아래와 같이 정의한다.

정의 1 (도메인): 도메인은 사용자 질의에 의해 결정되는 특정 과학기술 분야를 가리킨다. 사용자는 분석이 필요한 임의의 과학기술 분야를 질의로 입력할 수 있다. 특정 도메인 D에서 관련 문서 $C_D = \{d_1, d_2, ..., d_k\}$들을 얻을 수 있고, 각 문서 d_j 마다 전체 핵심문구들을 추출할 수 있다. 특정 도메인 전체 말뭉치에서 추출된 핵심문구는 $K_D = \{k_1, k_2, ..., k_l\}$이고 여기에서 $k_i \in d_j$이다.

정의 2 (문제점): 문제점은 특허 문서 C_D에서 해결하려고 하는 문제들로서, "recognizing signal patterns"와 같은 핵심문구로 표현된다. 문제점 핵심문구 집합은 전체 핵심문구 K_D의 부분 집합이며 $P_D = \{p_1, p_2, ..., p_m\} \subset K_D$와 같이 나타낼 수 있다. 경우에 따라 도메인은 상위 레벨의 문제점과 일치할 수 있다 (예: 음성 인식).

정의 3 (해결 방법): 해결 방법은 어떠한 문제를 해결하는 방법론, 모델 혹은 접근 방법 등을 가리키고, "은닉 마르코프 모델(Hidden Markov Model)"과 같은 핵심문구로 표현된다. 집합 $S_D = \{s_1, s_2, ..., s_n\} \subset K_D$로 표시할 수 있다. 핵심문구에는 문제점 혹은 해결 방법을 나타내는 핵심문구 외에 그와 관련된 문구들도 포함되므로, 문제점과 해결 방법을 나타내는 핵심문구의 합은 전체 핵심문구 수 보다 작다($m + n \leq l$).

정의 4 (특정 기간, time span): 특허 말뭉치에 포함되어 있는 특허의 전체 출원시간은 $T_C = \{\tau_1, \tau_2, ..., \tau_0\}$이고, 특정 기간(time span) l은 $\tau_i \leq l \leq \tau_j$기간이며 여기에서 $\tau_i, \tau_j \in T_c$이다. 문제점과 해결방법으로 정의되는 기술(technology)은 특정 시간정보를 갖는 특허 문서를 통해 추출되므로 그 기술이 출현하는 기간 정보가 중요한 역할을 한다. 상세한 내용은 3.3장에서 설명한다.

정의 5 (기술): 기술(technology)은 문제점, 해결 방법 및 도메인의 조합으로 정의하며, $t = <p_i s_j, D, \tau>$로 표시된다. 여기에서 p_i는 $p_i \in P_D, s_j \in S_D$이다. (예: "recognizing signal patterns" "hidden markov model" in "speech recognition").

정의 6 (과학기술 트렌드): 과학기술 트렌드(technology trend)는 특정 기간 내의 과학기술의 변화를 가리키며, 과학기술들간의 의미적 연관관계로서 표현된다. 연관관계는 동일한 문제점을 갖고 있는 과학기술 간의 연관관계(연관관계 1)와 동일한 해결 방법을 적용한 과학기술 간의 연관관계(연관관계 2)로 정의한다.

본 연구에서는 이러한 연관관계를 통해 과학기술의 트렌드를 탐지하는 것을 TTD라 정의하였다. 그림 2에 근거하면 (1) 기술 1과 기술 2는 같은 문제점 A을 해결하지만 특정 기간 l_1에서는 해결 방법 A로, 특정 기간 l_2에서의 해결 방법 B로 해결하였다. 따라서 기술 1이 기술 2로 발전되었다고 볼 수 있다; (2) 기술 4와 기술 5는 서로 다른 문제점을 동일한 해결 방법 C에 의해 해결하기에 두 기술은 연관성이 많다고 볼 수 있다; (3) 동일한 특정 기간 l_3에서 기술 3과 기술 6은 모두 해결 방법 E에 의해 서로 다른 문제점 E와 문제점 F가 해결되었다. 따라서 해결방법 E가 여러 분야에서 쓰이는 효과적인 해결 방법일 것이라는 추측이 가능하다.

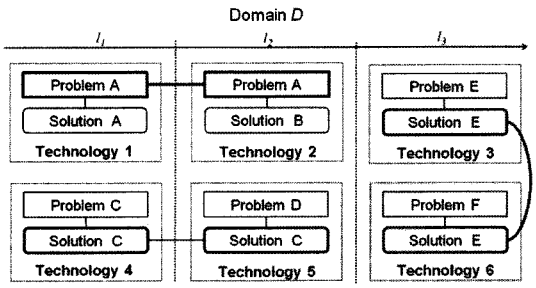

그림 2 과학기술 트렌드 탐지 예제

3. 방법론

본 연구의 목적은 과학기술의 트렌드를 탐지하는 것이다. 따라서 주요한 과학기술을 탐지하는 태스크 1과 이러한 과학기술 간의 의미적인 연관관계를 이용해 트렌드를 탐지하는 태스크 2를 수행해야 한다.

3.1 특허 정보

본 연구에서는 특허 데이터를 대표하는[9] 미국 특허청 USTPO[1]사이트에서 400개의 음성 인식 도메인 관련 특허 문서를 수집하고 문제점과 해결 방법에 대해 수동으로 태깅한 것을 실험에 사용하였다. 태깅 과정에

1) USPTO 사이트 http://www.uspto.gov/main/search.html

서 발견된 약어(acronym)(예: HMM) 관련 문제점은 특허 말뭉치와 Wikipedia[2]를 이용하여 구축한 약어 사전을 이용하여 보완하였다. 또한, 사전을 구축함에 있어서 WordNet[3]을 동시 적용하여 어휘적 변화(명사의 단수, 복수형; 동사의 변형형태)에 대해 정규화(normalize)하였다.

특허 문서는 구조적 항목(예: 출원날짜, 특허분류 등)과 비구조적 항목(예: 발명 명칭, 초록, 청구항 등)으로 나뉜다. 본 연구에서는 구조적 항목만 응용한 기존 연구[6-8]와 달리 두 가지 항목 모두를 이용하여 과학기술 트렌드를 탐지한다.

표 1에서 날짜는 특허 출원 날짜와 등록 날짜를 가리키는데 본 연구의 경우, 출원 날짜(Filed Date)를 타임스탬프로 정하였다. 인용관계(Referenced By) 필드는 인용된 다른 특허들을 가리킨다. 본 연구에서는 동일한 문제점 또는 해결 방법을 갖고 있는 특허들이 서로 인용할 가능성이 많다는 가정 하에 인용정보를 의미 핵심문구를 추출하는데 사용하였다. 청구항(Claims)은 특허의 핵심내용을 포괄하는 부분이고 여러 개의 청구항들로 구성되었다. 청구항(Claims)과 상세 설명(Description) 필드는 정규적인 텍스트로 구성되어 있기에 [9]에서 제시한 것과 같이 간단한 문법과 같은 언어적 특성을 이용한 핵심문구 추출이 가능하다. 나머지 필드정보(발명자, 출원인 등)는 본 연구의 고려 대상이 아니므로 이 부분에 대한 해석은 생략한다.

표 1 미국 특허 문서 구조

필드 명	값	필드 명	값
미국 특허 번호 (US Patent No.)	Number	국제 특허 분류 (International Patent Class)	Number
발명 명칭 (Title)	Free Text	인용관계 (Referenced By)	Patent Number
초록 (Abstract)	Free Text	**청구항(Claims)**	
발명자 (Inventors)	Proper Noun	**청구항 1 (Claim 1)**	Formulaic Free Text
출원인 (Assignee)	Proper Noun	**상세 설명(Description)**	
출원번호 (Application No.)	Number	**발명 배경 설명 (Background)**	Formulaic Free Text
출원 날짜 (Filed Date)	Date	**요약 (Summary)**	Free Text
등록 날짜 (Issue Date)	Date	**도면 설명 (Description of the Drawing)**	Free Text
미국 특허분류 (US Patent Class)	Number	**상세 설명 (Detailed Description)**	Free Text

2) Free Encyclopedia http://en.wikipedia.org/wiki/Main_Page
3) Lexical data base http://wordnet.princeton.edu/

3.2 의미 핵심문구 추출

전체 말뭉치 $C_D = \{d_1, d_2, ..., d_k\}$에서 문제점과 해결 방법을 포함한 의미 핵심문구 추출 과정(태스크1)은 아래와 같이 3개 단계로 나뉜다.

1 단계: 각 문서에서 모든 후보 핵심문구를 추출한다. 여기에서 핵심문구는 확률적 구문 구조 분석기[11]를 사용하여 탐색된 최단 명사구(noun phrase)와 명사구와 같은 레벨에 있는 동사구(verb phrase)까지를 포함하는 동사구(예: "recognize signal patterns")이다.

2 단계: 후보 핵심문구가 생성되면 분류기(classifier)를 이용하여 문제점 핵심문구를 추출하고, 이 문구 $P = \{p_{d_1}, p_{d_2}, \cdots, p_{d_n}\}$는 해결 방법 핵심문구 추출하기 위한 어휘적인 단서(lexical indicator)로 쓰인다.

3 단계: 나머지 후보 핵심문구에서 해결 방법 핵심문구 $S = \{s_{d_1}, s_{d_2}, \cdots, s_{d_n}\}$를 추출하여 최종적으로 과학기술 문구 집합 $T = \{t_{d_1}, t_{d_2}, \cdots, t_{d_n}\}$을 얻는다. 여기에서 t_{d_i}는 $t_{d_i} = (p_{d_k}, s_{d_l}, D, \tau_{d_j})$를 가리킨다(정의 4).

3.2.1 문제점 핵심문구 추출

본 연구에서는 문제점을 탐지할 수 있는 언어적 단서와 확률 모델을 결합하는 방법으로 문제점 핵심문구를 추출하였다. 문제점 핵심문구가 말뭉치 내에서 자주 발생한다는 토픽 키워드(topic keyword)의 특성에 기반한 언어 모델(language model)을 적용하였다. 예를 들면 문제점 핵심문구 "pattern recognition"이 문서와 말뭉치 내에서의 빈도가 높을 것이라고 예상할 수 있으나, 데이터 희소성(sparseness) 때문에 언어 모델을 적용하기가 쉽지 않고, 정보검색(information retrieval)에서 쓰는 평활법(smoothing)도 본 연구에서 적용하기 적합하지 않다. 따라서 후보 핵심문구 k의 확률은 아래와 같이 전체 단어 w_i(한 음절 모델)의 확률을 합하는 방법으로 추정 하였다.

$$p(k|d) = p(w_i|d) \sum_{i=1}^{|k|} p(w_i, d) \quad (1)$$

또한 단어 발생여부가 문맥에 의존하기 때문에 한 음절 모델(식 (1))을 두 음절 모델(식 (2))로 확장하였다.

$$p(k|d) = p(w_i|d) \sum_{i=2}^{|k|} p(w_i|w_{i-1}, d) \quad (2)$$

세 음절 이상의 모델은 데이터 희소성 때문에 의존관계를 고려하기 어렵고, 평활법을 적용하기에도 어려운 점이 있기에 본 연구에서는 한 음절 모델(uni-gram)과 두 음절 모델(bi-gram)만 고려하였다.

본 연구에서 인용문서 언어 모델과 배경 언어 모델(전체 말뭉치 기반 모델)을 결합한 혼합 모델을 사용하였다. 3.1장에서 제시하듯이 특허 문서는 인용정보를 포함하고 있을 뿐만 아니라 대부분의 문제점 핵심문구들은 인용된 문서와 본 문서에서 동시에 나타나고 주로 특허 문서의 발명 배경 설명 필드와 초록 필드에서 자주 나타난다. 따라서 본 연구에서는 문제점 핵심문구를 추출함에 있어서 해당 핵심문구가 인용문서와 전체 말뭉치에서 자주 나타날 것이라고 가정하였다. 문서 d에서 각 단어의 확률은 아래와 같이 계산된다. 여기에서 θ_R를 인용문서들에서의 단어 분포, θ_B는 배경 언어 모델, λ는 혼합 가중치(mixing weight)이다. 그리고 첫 항은 한 음절 모델, 둘째 항은 두 음절 모델이다.

$$p(w|d) = \lambda p(w|\theta_B) + (1-\lambda)\, p(w|\theta_R)$$

$$p(w_i|w_{i-1}, d) = \lambda\{p(w_i|w_{i-1}, \theta_B) \cdot p(w_{i-1}|\theta_B) + (1-\lambda)p(w_i|w_{i-1}, \theta_R) \cdot p(w_{i-1}|\theta_B)\} \quad (3)$$

θ_B와 θ_R에서의 단어의 확률은 다음과 같이 추정된다.

$$p(w|\theta_B) = \frac{\sum_{d_i \in C} c(w: d_i)}{\sum_{w' \in V_U} \sum_{d_i \in C} c(w': d_i)}$$

$$p(w|\theta_R) = \frac{\sum_{d_i \in R_d} c(w': d_i)}{\sum_{w' \in V_{UR}} \sum_{d_i \in R_d} c(w': d_i)} \quad (4)$$

여기에서 C는 말뭉치, R_d는 타겟 문서 d를 인용한 문서의 집합, V_U는 말뭉치 C 내의 한 음절 단어들의 집합, V_{UR}은 R_d중 한 음절 단어들의 집합, $c(w:d_i)$는 문서 d_i에서의 한 음절 단어 빈도수이다. 두 음절 모델에서의 단어 확률은 아래와 같이 추정된다.

$$p(w_i|w_{i-1}, \theta_B) = \frac{\sum_{d_i \in C} c(w_i|w_{i-1}: d)}{\sum_{w'_{i-1} w'_i \in V_B} \sum_{d_i \in C} c(w'_i|w'_{i-1}: d)}$$

$$p(w_i|w_{i-1}, \theta_R) = \frac{\sum_{d_i \in R_d} c(w': d_i)}{\sum_{w'_{i-1} w'_i \in V_{BR}} \sum_{d_i \in R_d} c(w'_i|w'_{i-1}: d)} \quad (5)$$

여기에서 $w_{i-1}w_i$는 두 음절, V_B는 말뭉치 C 내의 두 음절 단어들의 집합, V_{BR}은 R_d중 두 음절 단어들의 집합, $c(w_i|w_{i-1}:d)$는 문서 d_i에서의 두 음절 단어 빈도수이다. 이 외에 불용어(stop words)(예: a, the is)는 제거되었다. 하지만 배경 언어모델 θ_B는 주제 관련 특성(topicality)을 많이 표현하므로 이 모델만으로는 문제점 핵심문구를 추출하는데 부족하였다. 따라서 이를 보완하기 위해 인용문서 언어 모델 θ_R을 적용하였으나, 이것만으로 높은 성능을 기대하기는 어려웠다. 그리하여 주제 관련 특성을 제외한 문서에서 나오는 문제점 핵심문구 언어적 단서를 이용하는 방법을 추가하여 성능을 향상시켰다. 언어적 단서들은 학습 데이터에서 아래와 같은 이론적 근거 하에 추출되었다. 첫째, 전체 문제점 핵심문구의 57%에 해당되는 문제점 핵심문구는 특허 발명 명칭과 상세 설명 필드에서 나타나고(위치 정보), 둘째, 3.1장에서 설명한 바와 같이 일부 정규화된 형식을 쓰여진 필드에서 문제점 핵심문구와 관련된 패턴을 찾을 수 있다. 예를 들면, "method"와 "apparatus"는 문제점 핵심문구 앞에서 자주 나타난다. 학습 데이터에서 추출된 패턴들은 공통으로 포함하는 구문 정보를 이용하여 일반화하였는데 이 과정에서는 먼저 표면적 패턴(surface pattern)을 수집하고, 구문 구조 분석기를 통해 동일한 구문 정보를 가진 단어들을 하나로 합쳐서 일반화시켰다(예: method/*NN+PP*와 system/*NN+PP*은 (method|system)/*NN+PP*로 일반화).

표 2는 일반화된 패턴 샘플들이다. 위의 1단계에서 제시한 것과 같이 후보 문구는 명사구 또는 동사구이다.

표 2 문제점 핵심문구 패턴 샘플

No.	Lexico-Syntactic 패턴
1	method | apparatus | system | device + for | of + <Problem : NP | (VBG+NP)>
2	method | apparatus | system | device + <Problem : VP>
3	method | apparatus | system | device + WHNP + <Problem : VP>
4	to | of + <Problem : NP | (VBG+NP)> + with
5	for | of + <Problem : NP | (VBG+NP) > + using
6	to + <Problem : VP>

패턴 1과 2는 "system for verifying speakers"와 "device recognizes signals"와 같은 문장과 일치하고 패턴 3은 "system which removes noises"와 같은 문장들과 매칭된다. 패턴 4와 5는 "of verifying utterances with"와 "for noise reduction using"과 같은 문장들과 매칭되며 패턴 6은 "to summarize speech without decoding"과 매칭된다.

본 연구에서는 언어적 단서와 언어 모델을 결합하기 위하여 확률적 학습기를 이용하여 후보 문제 핵심문구를 분류하는 방법을 사용하였으며 패턴마다 가중치를 부여하여 분류기의 자질로 이용함으로써 성능을 높였다. 자질 함수(feature function) f는 아래와 같다.

$$f(k, \hat{p}_j: d_k) = \sum_{j=1}^{N} \delta(k, \hat{p}_j|d_k) \cdot \omega(\hat{p}_j)$$

$$\text{s.t.}\ \omega(\hat{p}_j) = \frac{\sum_{p_i \in training}(p_i, \hat{p}_j)}{\sum_{k_l \in training}(k_l, \hat{p}_j)} \quad (6)$$

여기에서 N은 전체 패턴 수이고, 핵심문구가 문서 d_k에서 패턴 $\hat{p}_j$에 의해 추출되었을 때 $\delta(k_l, \hat{p}_j \mid d_k)$의 값은 1이고 아니면 0이다. $\omega(\hat{p}_j)$는 패턴의 신뢰도로 학습 데이터에서 패턴 $\hat{p}_j$에 의해 추출된 정확한 문제점 핵심문구(p_i) 수가 차지하는 비율이다.

문제점 핵심문구를 추출하는 데는 다음과 같은 자질들이 사용되었다: (1) 한 음절 언어 모델(식 (1)), (2) 두 음절 언어 모델(식 (2)), (3) 필드 정보(표 1에서 볼드 체로 표시된 필드), (4) 언어적 패턴(linguistic pattern)(표 2), (5) 자질 함수(feature function)(식 (6)). 그 중에서 자질(1)과 (2)는 확률적 자질이고 나머지는 언어적 자질이다. 또한 식 (1)과 (2)의 자질 벡터에는 후보 핵심문구의 길이, 단어 확률과 후보 핵심문구의 확률을 포함된다. 이러한 자질들이 선형적으로 결합되어 전체 자질 공간을 이루게 된다.

3.2.2 해결 방법 핵심문구 추출

이 과정에서도 분류기를 이용하여 후보 핵심문구 중 해결 방법 핵심문구를 추출하였다. 해결 방법 핵심문구는 문제점 핵심문구와 달리 인용문서들 내에서 자주 나타나지 않으므로 언어 모델을 적용하기에 적합하지 않다. 따라서 주로 언어적 단서를 이용하여 그 핵심문구를 추출하였다. 태깅된 특허 문서 분석 결과를 보면 해결 방법 핵심문구는 문제점 핵심문구와 같이 자주 나타났다(예: "speech recognition(문제점) *using* language model(해결 방법)"). 그러므로 언어적 단서와 후보 핵심문구와 문제점 핵심문구의 동시 발생(co-occurrence) 정보를 이용해 주요 자질들을 설계할 수 있다.

태깅된 데이터에서 추출한 표면적 패턴(surface pattern)을 일반화하기 위해 문제점 핵심문구 추출에서와 같은 구문 정보를 이용하였다. 표 3은 일반화된 패턴 샘플들이다. 패턴 1은 "speech recognition using dynamic programming"과 같은 문장과 매칭되고, 패턴 2는 문장 "speaker verification by dynamic time warp"과, 패턴 3과 4는 각각 "linear discriminant analysis for speaker verification"과 "language model to recognize spoken language"와 매칭되며, 패턴 5는 "to assemble two acoustic samples"와 매칭된다. 자질

표 3 해결 방법 패턴 샘플

No.	Lexico-Syntactic 패턴
1	<Problem> + using | utilizing | employing + <Solution : NP | (VBG+NP)>
2	<Problem> + by | with + <Solution : NP | (VBG+NP)>
3	<Solution : NP | (VBG+NP)> + for | in + <Problem>
4	<Solution : NP | (VBG+NP)> + TO + <Problem : VP>
5	to + <Solution : VP>

함수(feature function) f는 아래와 같다.

$$f(k,\hat{p_j};d_k)=\sum_{j=1,\exists p_n\in P}^{N}\delta(k,p_n,\hat{p_j}|d_k)\cdot\omega(\hat{p_j})$$

$$\text{s.t. }\omega(\hat{p_j})=\frac{\sum_{p_m,s_i\in training}(s_i,p_m,\hat{p_j})}{\sum_{p_m,k_l\in training}(k_l,p_m,\hat{p_j})} \quad (7)$$

여기에서 핵심문구가 문서 d_k에서 문제점 핵심문구 집합 P내의 임의의 문제점 핵심문구(p_n)를 포함하고 있는 패턴 $\hat{p_j}$에 의해 추출되었을 때 $\delta(k_i,p_m,\hat{p_j}\mid d_k)$의 값은 1이거나 아니면 0이다. N은 전체 패턴 수이고, $\omega(\hat{p_j})$는 패턴의 신뢰도이며 이는 학습 데이터 중 문제점 핵심문구(p_m)와 같이 나타난 문구 가운데서 패턴 $\hat{p_j}$에 의해 추출된 정확한 해결 방법 핵심문구(s_i)가 차지하는 비율이다.

해결 방법 핵심문구를 추출함에 있어서는 언어적 패턴 외에 다른 자질들도 포함된다. 특허 문서 분석을 해보면 많은 해결 방법 핵심문구들이 "model"과 같은 키워드들을 포함하고 있음을 발견할 수 있다(예: "은닉 마코브 모델(Hidden Markov Model)", "언어 모델(language model)"). 이러한 키워드는 해결 방법 핵심문구를 추출하는 중요한 단서가 된다. 이것들을 본 연구에서 중심어(head word)라고 정의하였다. 중심어는 해결 방법 핵심문구에 포함되어 패턴을 적용하기가 어려우므로 분류기의 자질로만 사용된다. 자질로 사용된 중심어는 모두 11개이다("model, approach, method, methodology, technique, algorithm, analysis, measure, measurement, transform, structure"). 최종적으로 해결 방법 추출은 (1) 표 3의 패턴, (2) 자질 함수(feature function)(식 (7)), (3) 문제점 핵심문구와 제일 가까운 단어, (4) 11개의 중심어, (5) 구문 정보에 의해 학습되었다. 해결 방법 핵심문구 추출에서도 문제점 핵심문구 추출에서와 마찬가지로 선형적으로 기술된 자질들이 전체 자질 공간 벡터에 결합되게 된다.

3.3 과학기술 트렌드 탐지

본 장에서는 의미 핵심문구 추출 결과를 이용하여 과학기술을(정의 4) 추출하고 특정 기간(정의 6) 내의 핵심과학기술을 탐지하여 그들 간의 연관관계를 정의함으로써 과학기술 트렌드를 탐지하는 방법론을 소개한다.

정의 5에서처럼 타임 스탬프 정보를 가지고 특정 기간을 정의할 수 있는데, 그 중에서 가장 효과적인 특정 기간을 찾아내는 것이 본 연구의 태스크이다. 두 특정 기간에서의 언어 모델은 다르다. 즉 해결 방법의 변화 혹은 새로운 문제점의 출현과 같은 과학기술의 진보는 단어 분포에 변화를 가져다 주며 과학기술 트렌드를 탐지하는 기준이라고 볼 수 있다. 본 연구에서는 KL-divergence를 이용하여 두 언어 모델의 서로 다른 기간에서의 변화를 비교하였다.

$$D_{KL}(\theta_{t_2}\parallel\theta_{t_1})=\sum_{i=1}^{|V_{t_1}\cup V_{t_2}|}p(w_i|\theta_{t_2})\log\frac{p(w_i|\theta_{t_1})}{p(w_i|\theta_{t_2})} \quad (8)$$

여기에서 V_{l_1}는 특정 기간 l_1에 속하여 있는 전체 문서 내 단어들의 집합이다. KL-divergence가 비대칭적이므로 특정 기간 l_1와 특정 기간 l_2의 단어 분포의 차이를 비교함에 있어서 l_1이 l_2보다 먼저 발생하여야 한다. 특정 기간을 선택한 다음에는 그 기간 동안 핵심과학기술을 찾아야 하며, 과학기술의 중요성은 그 과학기술과 연관된 문서의 수로 계산된다. 이는 핵심과학기술일수록 많은 특허 문서가 인용되었다는 가정에 기반하여 아래와 같이 계산된다.

$$\text{importance}(t,l) = \frac{dc(t,l)}{dc(p_t,l)\cdot dc(s_t,l)} \quad (9)$$

여기에서 $dc(t,l)$는 특정 기간 l 내에 과학기술 t를 적용한 문서 수, t에 속하는 p_t는 문제점 핵심문구, s_t는 해결 방법 핵심문구다. 과학기술의 중요성은 그를 이루고 있는 문제점 핵심문구와 해결 방법 핵심문구를 포함하고 있는 문서수의 상호 정보량(mutual information)으로 계산하는데, 그 의미는 과학기술을 이루고 있는 문제점과 해결 방법 모두가 중요하다고 판단되었을 때 비로서 그 과학기술이 가장 중요하다는 것이다. 핵심과학기술을 탐지하는 과정은 다음과 같다.

과정 1: 특정 기간의 초기값은 말뭉치 내의 문서 수에 의해 정해진다.

과정 2: 모든 가능한 특정 기간의 조합으로 구성되며, 조합은 두 특정 기간이 연속적일 때만 유효하다. 특정 기간은 겹칠 수 있다(예: <1998~2000, 1999~2001>).

과정 3: 과정 2에서 나온 모든 조합 쌍의 KL-divergence를 계산하고 결과를 순위화한다.

과정 4: 과정 3에서 결정된 상위 쌍들 중에서 핵심과학기술들을 선택한다.

핵심과학기술들을 추출하면 정의 6에서 제시한 것과 같이 추출된 과학기술 간의 의미적 연관관계를 정의할 수 있다. (만약 두 기술 A와 B가 서로 다른 기간에서 동일한 문제점을 서로 다른 해결 방법에 의해 해결하였을 경우, 기술 A는 기술 B로 발전하였다고 볼 수 있다).

4. 실험결과

본 연구의 평가과정에서는 음성 인식 도메인에서의 과학기술 트렌드를 찾기 위한 시나리오를 디자인하여 사용자가 TTD 시스템을 통한 효과적인 정보 획득 방법에 대해 검증하였다.

4.1 실험 설정

실험에서는 1976년부터 2003년까지의 음성 인식 도메인 특허 1,420개를 사용하였고 컴퓨터 전공 대학원생 3명이 400개 샘플 데이터에 대해 의미 핵심문구(즉, 문제점과 해결 방법)을 태깅하여 평가 집합을 만들었다. 샘플 데이터는 전체 말뭉치 데이터 분포에 근거하여 선택하였고 최종 평가 집합은 각 특허마다 문제점 혹은 해결 방법 핵심문구에 대해 태깅 되었는데 이는 2명 이상의 동의에 의해 결정되었다. 전체 400개 문서 중 78% (300개 특허 문서)가 최종 평가 데이터(gold standard)로 결정되었으며, 이는 334개의 문제점 핵심문구와 311개 해결 방법 핵심문구로 구성되어, 태스크 1의 평가 기준으로 쓰였다. 태스크 2(핵심과학기술 탐지)의 평가는 태스크 1처럼 엄밀하지 않다. 많은 특정 기간 정보와 그 기간에 속하여 있는 대량의 문서들 때문에 태스크 1과 같이 엄밀한 평가를 하는 것이 본 연구에서는 불가능하였다.

4.2 평가

태스크 1은 정확률과 재현률을 평가 방법으로 정하였고, 태스크 2에 대해서는 탐지된 과학기술 트렌드를 분석하는 방법으로 평가를 진행하였다.

3.2절에서 제시한 혼합 언어 모델(mixture language model)에서 배경 언어 모델의 혼합 가중치는 λ=0.28로 경험적으로 설정하였고, 기계학습기는 LIBSVM[18]을 사용하였다. 문제점과 해결 방법 핵심문구 추출 태스크는 3중 검증(3-fold validation) 방법으로 실험을 진행하였으며 매번 200개의 데이터로 SVM 분류기를 학습시키고 100개의 데이터로 테스트를 수행하였다. 매번 200개 문서의 6,310개 정도의 후보 핵심문구가 학습됨으로 학습 데이터로 사용하기에 충분하였다. 평가방법은 R-정확률(즉, 각 재현률 시점에서의 평균 정확률)을 사용하였는데 이는 일반적으로 특허 문서에는 하나의 문제점과 해결 방법이 존재하지만 다수의 문제점과 해결 방법이 존재하는 특허 문서들도 발견되었기 때문이다.

문제점 핵심문구 추출 실험에서는 다음과 같은 두 가지 가정에 대해 실험하였다; (1) 3.2.1장에서 제시한 언어적 단서들의 유용성; (2) 패턴 가중치가 SVM 분류기에 주는 영향. 표 4는 언어 모델을 한 음절 모델과 두 음절 모델 2가지로 실험한 결과이다. 3.2.1장에 제시한 것과 같이 희소성 문제가 두 음절 모델을 이용한 결과에 영향을 미쳤으나 언어적 단서들에 의해 결과가 많이 보완되었다. 또한 언어적 패턴에 포함된 필드 정보(표 1)도 결과 향상에 영향을 주었으며, 패턴 가중치를 적용한 것이 적용하지 않은 것보다 성능을 향상시켰다.

표 4 문제점 핵심문구 추출 결과

자질		R-정확률
언어 모델	두 음절 모델	0.38 (-34%)
	한 음절 모델	0.58 (0%)
언어 모델+ 언어적 단서	패턴가중치 미적용 (수식 6 적용)	0.71 (+22%)
	패턴가중치 적용 (수식 6 적용)	0.76 (+31%)

해결 방법 핵심문구 추출은 3.2.2에서 제시한 자질을 가지고 평가를 진행하였다. 언어적 패턴만 사용한 경우와 다른 자질을 추가한 다음의 실험 결과는 표 5와 같다. 표 5에 따르면 중심어를 사용하였을 때의 결과가 가장 많은 향상이 있었음을 알 수 있다. 이는 대부분 해결 방법 핵심문구가 중심어를 포함하고 있기 때문이다. 패턴 가중치는 문제점 추출 과정에서보다 유력하지 않았지만 결과를 향상시켰으며 해결 방법과 문제점 핵심문구 간 거리 정보도 결과에 영향을 주었다.

표 6에서 보면 문제점 핵심문구 추출 모델이 "voice recognition"과 같은 넓은 의미의 문제점뿐만 아니라 "reduce the storage space for the speech recognition dictionary"과 같은 좁은 의미의 문제점 핵심문구도 추출함을 알 수 있다. 하지만 좁은 의미의 문제점들은 너무 구체적으로 표현되었기에 과학기술 트렌드를 표현하기에는 큰 의미는 없다. 또한 "speech recognition"이 말뭉치 내에서의 빈도수가 높아 결과가 "speech recognition"에 치중되었다. 유사어(synonym) 문제도 존재하는데 이는 두 문구가 표면적(surface) 레벨에서는 서로 다르게 쓰였지만 의미적으로 같은 경우이다. 예를 든다면 "speaker recognition"과 "voice identification"의 경우 의미상으로 거의 비슷하지만 이 문제는 WordNet의 유사어 정보로도 해결되기 어렵다.

표 5 해결 방법 핵심문구 추출 결과

자질공간 (Feature Space)	R-정확률
언어적 패턴	0.62 (0%)
+ 패턴 가중치 (수식 7)	0.66 (7%)
+ 11 중심어(head words)	0.74 (19%)
+ 해결 방법과 문제점 핵심문구간 거리	0.75 (21%)

표 6 의미 핵심문구 추출 샘플

문제점 핵심문구
speech recognition, pattern recognition, noise reduction, reduce storage space for speech recognition, , voice identification, speaker recognition, recognition error reduction
해결 방법 핵심문구
dynamic programming, vector quantization method, shared speech model, lattice-ladder filters, user-cued speech recognition, acoustic model, neural network, dynamic programming algorithm

해결 방법 추출 결과를 보면 "dynamic programming", "hidden markov model", "language model"과 같은 명확한 해결 방법 핵심문구가 있는 반면에 다양한 형식의 결과를 포함하고 있다.

의미 핵심문구 추출 결과에서 보면, 대부분 오류는 시스템이 너무 서술형으로 표현된 문구들(예: "transform the consistent message into electrical signal representation and generate a likelihood score of recognition")을 추출하는 데서 발생되었다. 특히 해결 방법 핵심문구는 서술형으로 표현되는 경우가 많은데, 그것들이 실험한 결과에서 제대로 추출되지 않았다. 하지만 긴 문구들은 너무 구체적인 표현이기에 본 연구에서 과학기술 트렌드를 탐지하는데 중요하지 않다고 판단하여 의미 핵심문구 추출 결과에는 과학기술 분석에 유용한 짧은 문구들만을 포함시켰다.

태스크 2에서는 3.3절에서 제시한 4단계 과정을 거쳐 과학기술 트렌드가 탐지되었다. 실험에서 특정 기간의 초기값은 1년으로 하고 1,420개 특허로 비지도 학습 방법으로 실험을 진행하였다. 표 7은 TTD 실험 결과 일부를 나타낸 것이다.

표 7에서 보듯이 핵심과학기술이 "speech recognition"이라는 문제점과 관련이 있다는 것을 알 수 있고, 1999년부터 "speaker verification"이라는 새로운 문제점이 존재하기 시작하였음을 탐지할 수 있다. 그리고 "speech recognition" 해결 방법은 1980년의 "dynamic programming"에서 1990년대에는 "hidden markov model"과 "language model"로 변화되었음을 알 수 있다. "dynamic time warping "은 1980년대 나왔다가

표 7 과학기술 트렌드 탐지 결과 샘플

특정 기간	1980-1983	1980-1982	1986-1987	1988-1989	1992-1996
과학기술	**speech recognition**	**speech recognition**	**speaker verification**	**pattern matching**	**speech recognition**
	dynamic programming	*dynamic time warping*	*dynamic time warping*	*dynamic programming*	*dynamic time warping*
특정 기간	1994-1996	1997-1998	1998-2001	1999 2000	2000
과학기술	**speech recognition**	**speech recognition**	**speech recognition**	**speaker verification**	**language recognition**
	hidden markov model	*language model*	*hidden markov model*	*language model*	*language model*

1990년대에 다시 흥행하였는데 이는 1992년에 "dynamic time warping"방법론에 성능 향상에 필요한 새로운 자질들이 추가되었다고 추측할 수 있으나 본 연구에서는 탐지하기 어려웠다.

5. 관련 연구

특허 관련 기존 연구는 특허 검색 및 분류 와 특허 분석으로 나뉜다.

특허 검색 관련 연구는 효과적으로 특허를 검색하는 것을 목적으로 한다. [12]에서는 청구항 구조를 분석하여 검색 태스크의 성능을 높였고 [13]에서는 단어 분포를 이용하여 기술 검색 태스크 성능을 향상시켰고, [14]에서는 bag-of-word 방법론을 적용하여 특허 문서를 분류, [15]에서는 인용정보 분석을 통해 특허 분류 시스템 성능을 향상시켰다.

특허 분석 관련 연구는 특허 문서 분석을 통하여 의미 있는 정보를 추출하여 여러 분야에 응용하는 것이다. [3]에서는 시간에 따른 문구 빈도수의 변화를 Shape Query Language로 표현함으로써 특허 말뭉치에서 기술 트렌드를 가시화시켰고, [6]에서는 계층적 클러스터링(hierarchical clustering) 기법[10]으로 클러스터링된 특허 문서에서 신생의 개념을 추출하였다. [7]에서는 특허망을 구축하여 기술의 발전으로 탐지하였고, [8]에서는 특허 지도(patent map)를 만들어 신생 과학기술을 탐지하여 가시화시켰고, [9]에서는 과학기술의 진화를 탐지하였으며, [5]는 특허 시스템을 구축하여 기술의 진보를 의미적으로 보여줌으로써 사용자에게 편리를 주었다. [16]에서는 수사학적 구조를 분석하는 방법을 적용하여 청구항을 쉽게 읽는 방법을 제안하였으며, [17]에서는 여러 조직의 기술정보 보급에 관한 연구를 시도하였다.

6. 결 론

특허 문서에서의 과학기술 트렌드 탐지는 최근 과학기술을 이해하고 특허 검색 결과에 의해 발생하는 정보 과잉 문제를 해결하는데 큰 도움을 준다. 본 연구에서는 특허 문서 내에서 과학기술을 나타내는 문제점 핵심문구와 해결 방법 핵심문구를 추출하고 과학기술 간의 의미적 연관관계를 통해 과학기술 트렌드를 자동으로 탐지하는 TTD 시스템을 제안함으로써 과학기술의 주요한 흐름을 쉽게 이해할 수 있도록 하였다. 이 시스템을 검증하기 위해 실행된 실험에서는 의미 핵심문구 추출 태스크에 대해 엄밀한 평가를 진행하였으며 그 결과에 근거하여 음성 인식 분야에서 의미 있는 과학기술 트렌드를 볼 수 있었다. 비록 실험에서는 도메인을 음성 인식 분야 한 가지로 지정하여 진행하였지만 의미 핵심문구 추출에 사용한 자질들은 다른 과학기술 도메인에서 적용이 가능하다.

향후 연구로 4.2절에서 언급한 유사어 문제 해결과 태스크 2의 평가 방법 보완을 목표로 진행하고자 한다.

참 고 문 헌

[1] Q.Mei and C.Zhai. A mixture model for contextual text mining. In Proceedings of the 12th ACM SIGKDD international conference on Knowledge Discovery and Data mining(KDD'06), pp. 649-655, 2006.

[2] R. Nallpati. Semantic language models for topic detection and tracking. In Proceedings of the conference of the North American chapter of the Association for Computational Linguistics on Human Language Technology (HLTNAACL'03), pp. 1-6, 2003.

[3] B. Lent, R. Agrawal, and R. Srikant. Discovering trends in text databases. In Proceedings of the 3rd international conference on Knowledge Discovery and Data mining (KDD'97), pp. 227-230, 1997.

[4] A. Porter and D. Jhu. Technological mapping for management of technology. In Proceedings of International Symposium on Technology, 2001.

[5] L. Wanner, et al. Towards content-oriented patent document processing. World Patent Information, Vol. 30 (1), pp. 21-33, 2007.

[6] W. Pottenger and T. Yang. Detecting emerging concepts in textual data mining. Computational Information Retrieval, pp. 1-17, 2001.

[7] B. Yoon and Y. Park. A text mining-based patent network: analytical tool for high-technology trend. Journal of High Technology Management Research, Vol. 15 (1), pp. 37-50, 2004.

[8] Y. Kim, J. Suh, and S. Park. Visualization of patent analysis for emerging technology. Expert Systems with Applications,Vol. 34 (3), pp. 1804-1812, 2007.

[9] K. Ahmad and A. Al-Thubaity. Can text analysis tell us something about technology progress? In Proceedings of the ACL-03 workshop on patent corpus processing, pp. 41-45, 2003.

[10] F. Bouskila and W. Pottenger. The role of semantic locality in hierarchical distributed dynamic indexing. In Proceedings of the International Conference on Artificial Intelligence (IC-AI'00), 2000.

[11] D. Klein and C. Manning. Accurate unlexicalized parsing. In Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics (ACL-03), pp. 423-430, 2003.

[12] T. Takaki, A. Fujii, and T. Ishikawa. Associative document retrieval by query subtopic analysis and

its application to invalidity patent search. In Proceedings of the 13th ACM International conference on Information and Knowledge Management (CIKM '04), pp. 399-406, 2004.
[13] H. Itoh, H. Mano, and Y. Ogawa, Term distillation in patent retrieval. In Proceedings of the ACL-03 workshop on patent corpus processing, pp. 41-45, 2003.
[14] C. Koster, M. Seutter and J. Beney. Multi-Classification of Patent Applications with winnow. In Proceedings PSI 2003, pp. 545-554, 2003.
[15] K. Lai and S. Wu. Using the patent co-citation approach to establish a new patent classification system. Information Processing and Management, Vol. 41, pp. 313-330, 2005.
[16] A. Shinmori, M. Okumura, Y. Marukawa, and M. Iwayama.Patent claim processing for readability: structure analysis and term explanation. In Proceedings of the ACL-03 workshop on patent corpus processing, pp. 56-65, 2003.
[17] A. Chakrabarti, I. Dror, and N. Eakabuse. Interorganizational transfer of knowledge: An analysis of patent citations of a defense firm. IEEE Transactions on Engineering Management, Vol. 40 (1), pp. 91-94, 1993.
[18] Library for Support Vector Machine http://www.csie.ntu.edu.tw/~cjlin/libsvm/

전 영 실
2005년 중국 연변과학기술대학 컴퓨터공학과(학사). 2006년~현재 한국과학기술원 정보통신공학과 석사과정. 관심분야는 정보검색 및 자연어처리, 텍스트마이닝 등

김 영 호
2006년 인하대학교 컴퓨터 공학부(학사) 2006년~현재 한국과학기술원 정보통신공학과 석사과정. 관심분야는 정보검색 및 자연어처리, 텍스트마이닝, 인공지능 등

정 윤 재
정보과학회논문지 : 소프트웨어 및 응용 제 36 권 제 3 호 참조

류 지 희
2006년 충남대학교 전기정보통신공학부(학사). 2007년~현재 한국과학기술원 정보통신공학과 석사과정. 관심분야는 정보검색 및 자연어처리, 데이터마이닝, 정보추출, 온톨로지 등

맹 성 현
정보과학회논문지 : 소프트웨어 및 응용 제 36 권 제 3 호 참조