지식경제부 산업융합원천기술개발사업

선행기술조사 보고서

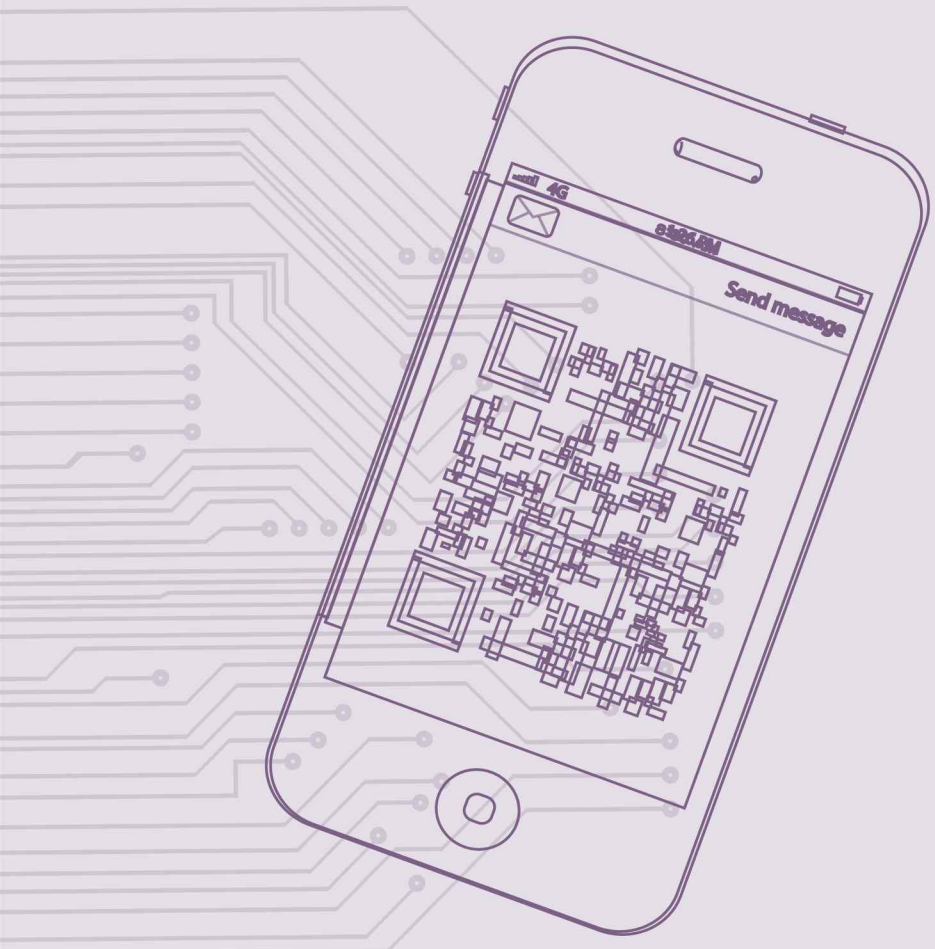
SW분야

글로벌 유통형 비정형 빅데이터 검색 SW / 산업현장의 니즈를 직접 반영한 빅데이터 분석·응용 SW

2013. 2

Keit 한국산업기술평가관리원

목차



1. 과제정보

과제관리번호	SW
과제명	글로벌 유통형 비정형 빅데이터 검색 SW 산업현장의 니즈를 직접 반영한 빅데이터 분석·응용 SW

2. 조사결과(주요 선행기술문헌기준)

■ 기술요소별 비교

기술요소		국내문헌	국외문헌
A	빅데이터	△	○
B	비정형데이터의 정형분석을 위한 Text Mining/Auto Labeling	○	○
C	다국어 동시 검색 및 각 언어별 검색기능	△	○

기술 요소	국내문헌 비교		
	한국등록특허 697,689 ((주)공영디비엠)	한국공개특허 2000-0049928 (한동윤)	국내문헌 기술별 종합유사도
A	△	X	△
B	○	X	○
C	X	△	△

기술 요소	국외문헌 비교		
	미국등록특허 8,027,994 (International Business Machines Corporation)	미국등록특허 7,512,602 (International Business Machines Corporation)	국외문헌 기술별 종합유사도
A	○	○	○
B	○	○	○
C	○	X	○

3. 조사의견

기술요소 A	**주요문헌**	■ 한국등록특허 697,689 ■ 미국특허 8,027,994 ■ 미국특허 7,512,602
유사점	문서를 포함하는 데이터베이스, 비정형적인 주소 또는 과거 주소정보 데이터 등 대용량 데이터베이스 기술 분야인 점	
차이점	-	
기술요소 B	**주요문헌**	■ 한국등록특허 697,689 ■ 미국특허 8,027,994 ■ 미국특허 7,512,602
유사점	비정형데이터인 데이터베이스 내의 문서로부터 키워드를 추출하고, 추출된 키워드에 해당하는 **타국어와의 색인을 생성하는 기술** 및 수집되는 비정형 문서 메타 데이터 정보를 정형분석을 위해 문서 라벨 분석 및 텍스트 분석을 수행하는 점	
차이점	-	
기술요소 C	**주요문헌**	■ 한국공개특허 2000-0049928 ■ 미국특허 8,027,994
유사점	데이터베이스 내의 문서로부터 추출된 키워드에 해당하는 타국어와의 색인을 생성하며, 생성된 색인을 통해 **자국어 검색 시 타국어 문서를 동시에 검색**할 수 있도록 한 점	
차이점	-	

※ 동 보고서는 연구제안서를 바탕으로 주요기술 중 기술이 명확하게 드러난 구성을 중심으로 선행특허조사한 결과임

4. 분석기준 및 분석방법

4-1. 조사대상

조사대상 국가	한국	미국	일본	EP	국제특허	기타
	○	○	○	○	○	
조사대상 기간	~ 2012. 12. 3 (조사개시일 이전 공개자료)					

4-2. 기술분류

IPC	· G06F 17/00, G06F 17/30, G06F 17/40, G06F 9/44

4-3. 검색방법

조사관점	개발 목표 및 내용을 참고하여 기술요소 A와 기술요소 B, 기술요소 C로 분류하였음	
	비정형 빅데이터의 검색 시스템과 비정형데이터의 정형분석을 위한 text mining/auto labeling기술, 다국어 동시 검색 기술을 중심으로 조사를 실시하였음	
키워드	국문	빅데이터, 하둡, 비정형, 형태소, 다국어
	영문	big data, hadoop, morph, multi alnguage, text mining, label
검색식	1. (multi* and language* and data*) and (unformat* or unstructur* or unconstruct* or atpical* or freeform* or (free* adj form*) or notype* or nontype* or uniform* or deformabl* or irregularity* or anomal*) and (search*).TI. 2. (다국어* or 다언어* or 다중어* or (다중* adj 언어*) or (multi* adj (lingual* or langage*)) and (search*).TI. 3. (빅데이터* or ((빅* or 대용량*) adj 데이터) or (big* adj data*)) and (비정형* or 불균일* or 불규칙* or unformatt* or unstructure* or atypical* or freeform* or (free* adj form*) or (non* adj fix*) or nonfigid* or deformabl* or irregular* or anomaly* or anomalist* or unfix*) and (mining* or label*) 4. (빅데이터* or ((빅* or 대용량*) adj 데이터) or (big* adj data*)) and (hadoop* or 하둡* or HDFS*)	

5. 주요 선행기술문헌

문헌번호	기술요지	기술요소	관련도
미국등록특허 8,027,994	문서를 포함하는 데이터베이스를 검색하는 방법으로서, 각 **문서로부터 키워드를 추출하고, 키워드 사전을 이용하여 키워드를 지원언어(타국어)로 번역하고, 지원언어로 된 문서에 대한 키워드 리스트를 생성하여 문서의 역색인을 생성하며, 자국 검색어로 검색 시, 자국 검색어 및 다른 지원언어(타국어)로 된 동의어를 이용하여 검색 문서를 식별하고, 검색된 문서의 리스트를 제공**하는 것을 특징으로 함	A B C	○ ○ ○
미국등록특허 7,512,602	수집되는 비정형 문서 메타 데이터 정보를 정형분석을 위해 문서 라벨 분석 및 텍스트 분석을 수행하는 것을 특징으로 함	A B C	○ ○ X
한국등록특허 697,689	**비정형적인 주소 또는 과거 주소정보 데이터를 문자열 정형화 시켜, 문자패턴을 분석하여 패턴에 의한 주소구성을 개별항목으로 분리하여 레퍼런스 정보와 매핑**하여 최신 우편번호를 찾기 위한 방법을 특징으로 함	A B C	△ ○ X
한국공개특허 2000-0049928	다국어를 동시에 검색하기 위한 검색엔진 기술을 특징으로 함	A B C	X X △

※ 선행특허의 소유권자, 공개일자 및 구체적인 기술적 내용은 「기술구성의 대비」를 참조

6. 기술구성의 대비

<table>
<tr><td colspan="2">일련번호 1</td><td colspan="3">미국등록특허 8,027,994</td></tr>
<tr><td colspan="2">출원일자</td><td>2008. 08. 21</td><td>등록일자</td><td>2011. 09. 27</td></tr>
<tr><td colspan="2">특허권자</td><td colspan="3">International Business Machines Corporation</td></tr>
<tr><td colspan="2">제목</td><td colspan="3">Searching a multi-lingual database</td></tr>
<tr><td colspan="5">구 성 대 비</td></tr>
<tr><td colspan="2">제안기술</td><td colspan="3">선행기술</td></tr>
<tr><td colspan="2">A. 빅데이터
– 빅데이터 저장구조인 Hadoop 데이터 수집기/색인 병렬 저장기

B. 비정형데이터의 정형분석을 위한 Text Mining/Auto Labeling

C. 다국어 동시 검색 및 각 언어별 검색기능
– 한, 중, 일, 영 형태소 분석기 개발로 각 언어별 검색기능
– 다국어 문서들의 자동 인식 색인 처리로 다국어 동시 검색</td><td colspan="3">□ 기술요지
○ 다수의 언어로 이루어진 자료를 포함하는 데이터베이스를 검색하기 위한 검색 시스템 및 방법

□ 제안기술 A, B, C 관련
○ (page 11, 청구항 1~3, 11, 16항]
○ 문서를 포함하는 데이터베이스를 검색하는 방법으로서, 각 문서로부터 키워드를 추출하고, 키워드 사전을 이용하여 키워드를 지원언어(타국어)로 번역하고, 지원언어로 된 문서에 대한 키워드 리스트를 생성하여 문서의 역색인을 생성하며, 자국 검색어로 검색 시, 자국 검색어 및 다른 지원언어(타국어)로 된 동의어를 이용하여 검색 문서를 식별하고, 검색된 문서의 리스트를 제공</td></tr>
<tr><td colspan="5">검 토 의 견</td></tr>
<tr><td colspan="5">제안기술과 관련문헌의 비교결과, 문서를 포함하는 데이터 베이스의 각 문서로부터 키워드를 추출하고, 추출된 키워드에 해당하는 타국어와의 색인을 생성하며, 생성된 색인을 통해 자국어 검색 시 타국어 문서를 동시에 검색할 수 있도록 한 점</td></tr>
</table>

일련번호	2	미국등록특허 7,512,602		
출원일자		2006. 11. 30	등록일자	2009. 03. 31
특허권자		International Business Machines Corporation		
제목		System, method and computer program product for performing unstructured information management and automatic text analysis, including a search operator functioning as a weighted and (WAND)		

구 성 대 비

제안기술	선행기술
A. 빅데이터 – 빅데이터 저장구조인 Hadoop 데이터 수집기/색인 병렬 저장기 B. 비정형데이터의 정형분석을 위한 Text Mining/Auto Labeling C. 다국어 동시 검색 및 각 언어별 검색기능 – 한, 중, 일, 영 형태소 분석기 개발로 각 언어별 검색기능 – 다국어 문서들의 자동 인식 색인 처리로 다국어 동시 검색	□ 기술요지 ○ 비정형 정보 관리와 자동 텍스트 분석을 수행하기 위한 컴퓨터 프로그램 □ 제안기술 A관련 ○ (page 43, 발명의 상세한 설명] ○ 문서 메타 데이터 □ 제안기술 B관련 ○ (page 38~41, 발명의 상세한 설명] ○ 비정형 정보의 정형분석을 위해 문서 라벨 분석 및 텍스트 분석을 수행

검 토 의 견

제안기술과 관련문헌의 비교결과, 다국어의 동시 검색 기술은 관련 문헌에 기재되지 않았으나, 수집되는 비정형 문서 메타 데이터 정보를 정형분석을 위해 문서 라벨 분석 및 텍스트 분석을 수행하는 점

일련번호	3	한국등록특허 697,689		
출원일자		2005. 08. 10	등록일자	2007. 03. 14
특허권자		(주)공영디비엠		
제목		비정형 데이터베이스의 정형화 장치를 이용한 정형화 방법		

구 성 대 비

제안기술	선행기술
A. 빅데이터 – 빅데이터 저장구조인 Hadoop 데이터 수집기/색인 병렬 저장기 B. 비정형데이터의 정형분석을 위한 Text Mining/Auto Labeling C. 다국어 동시 검색 및 각 언어별 검색기능 – 한, 중, 일, 영 형태소 분석기 개발로 각 언어별 검색기능 – 다국어 문서들의 자동 인식 색인 처리로 다국어 동시 검색	□ 기술요지 ○ 비정형의 데이터베이스를 정형화된 자료로 정리하기 위한 정형화 장치 및 정형화 방법 □ 제안기술 A관련 ○ (page 5, 발명의 상세한 설명] ○ 비정형적인 주소 또는 과거 주소정보 데이터 □ 제안기술 B관련 ○ (page 2, 청구항 3항] ○ 비정형 데이터베이스에 저장된 주소정보를 문자열 정형화 시키고, 단어단위로 패턴화하고, 저장된 패턴 유형에 따라 문자열 특성에 저장

검 토 의 견

제안기술과 관련문헌의 비교결과, 다국어의 동시 검색 기술은 관련 문헌에 기재되지 않았으나, 비정형 데이터 베이스를 정형화된 자료로 정리하기 위해 text mining을 이용하고, 비정형적인 주소 또는 과거 주소정보 데이터를 문자열 정형화 시켜, 문자패턴을 분석하여 패턴에 의한 주소구성을 개별항목으로 분리하여 레퍼런스 정보와 매핑하여 최신 우편번호를 찾기 위한 방법

일련번호	4	한국공개특허 2000-0049928		
출원일자		2000. 05. 08	공개일자	2000. 08. 05
특허권자		한동윤		
제목		다국어 검색엔진의 운영 장치 및 방법		

구 성 대 비

제안기술	선행기술
A. 빅데이터 – 빅데이터 저장구조인 Hadoop 데이터 수집기/색인 병렬 저장기 B. 비정형데이터의 정형분석을 위한 Text Mining/Auto Labeling C. 다국어 동시 검색 및 각 언어별 검색기능 – 한, 중, 일, 영 형태소 분석기 개발로 각 언어별 검색기능 – 다국어 문서들의 자동 인식 색인 처리로 다국어 동시 검색	□ 기술요지 ○ 외국어에 능숙하지 않은 인터넷 이용자가 한국어로 검색하면 외국의 검색 엔진을 통해 검색 결과를 한국어로 볼 수 있도록 하는 다국어 검색엔진의 운영 장치 및 방법 □ 제안기술 C 관련 ○ (page 3, 청구항 1~3항] ○ 한국어 키워드를 입력하여 검색 시 입력 키워드를 다국어로 번역하여 검색을 수행하고 검색 결과를 웹브라우저 상에 출력하여 다국어로 되어 있는 데이터를 동시에 검색하는 기술

검 토 의 견

제안기술과 관련문헌의 비교결과, 빅데이터 기술 및 비정형 데이터의 정형분석을 위한 Text Mining/Auto Labeling 기술은 기재되지 않음 다국어를 동시에 검색하기 위한 검색엔진 기술은 일부 기재됨

7. 참고 선행기술문헌 리스트

동 연구과제에 대하여 국내외 선행특허기술을 키워드와 국제특허분류를 이용하여 조사한 결과 해당 연구계획을 이해하는데 도움이 되는 기초자료

문헌번호	특허권자 (논문저자)	발명의 명칭
미국특허 8,306,972	Google Inc.	Ordering of search results based on language and/or country of the search results
미국특허 8,027,966	International Business Machines Corporation	Method and system for searching a multi-lingual database
미국특허 7,433,894	International Business Machines Corporation	Method and system for searching a multi-lingual database
미국특허 6,968,338	The United States of America as represented by the Administrator of the National Aeronautics and Space Administration	Extensible database framework for management of unstructured and semi-structured documents
미국특허 6,952,691	International Business Machines Corporation	Method and system for searching a multi-lingual database
미국공개특허 2009-0024599	Giovanni Tata	METHOD FOR MULTI-LINGUAL SEARCH AND DATA MINING
한국등록특허 1,158,864	동국대학교 경주캠퍼스 산학협력단	맵리듀스 기반의 대용량 데이터 분산 계산 방법 및 그 시스템(Distributed computation method and system based on mapreduce of large amount data)
한국등록특허 912,371	한국전자통신연구원	클러스터 환경에서 고확장성을 지원하는 대용량 고차원데이터 색인 장치 및 방법(Indexing System And Method For Data With High Demensionality In Cluster Environment)
한국등록특허 835,706	한국과학기술정보연구원	자동 색인을 위한 한국어 형태소 분석 시스템 및 그 방법(System and method for korean morphological analysis for automatic indexing)
한국공개특허 2001-0034973	세림정보기술 주식회사	전문검색과 웹 게시판 자료연동에 의한 데이터 구축 및검색 시스템(Construct and reference by FTR and peristalsis of data in the web bulletin board)
일본등록특허 3,455,641	TOSHIBA CORP	지식 정보 검색 시스템 및 지식 정보 검색 방법