마이닝기법을 활용한 정보분석 모델

2003년도 한국기술혁신학회
추계학술회의 및 콜로퀴엄 발표논문

2003. 11. 29
과학기술연구원 국제회의실

박 철 균 팀장 (아주대학교 전산.....)
배 상 진 팀장 (과학기술정보연구원 산업분석실)
정 용 일 (과학기술정보연구원 산업분석실)

마이닝 기법을 활용한 정보분석 모델의 가능성 연구

2003. 11

박철균, 배상진, 정용일

KISTI Korea Institute of Science and Technology Information

구 성 내 용

1. 연구배경
2. 정보분석 시스템화 사례조사
3. 정보분석시스템 모델 개발
4. 구성기술 고찰
5. 결론 및 향후 연구

연구배경

- 정보분석 환경의 변화
  - 과학기술문헌 발간량의 증대, 인터넷을 통한 정보 홍수 → 정보분석가의 정보입수에 대한 부담 가중
  - 정보분석의 객관성, 계량화 요구, 생산성 향상 요구 증대 → 새로운 정보분석 방법론 개발
  - 전산시스템에 의한 정보분석 관련 기술 발전 : 계량서지학, 마이닝, 정보추출(information extraction)
- 정보분석시스템의 최근 동향 : 정형필드만을 통계적, 계량서지학적으로 처리하는데 따르는 한계를 극복하기 위하여 초록과 같은 비정형 필드를 분석 대상 필드로 확장하려는 움직임을 보이고 있으며, 이를 위하여 텍스트마이닝과 같은 기술의 적용이 시도되고 있음.
- 웹텍스트와 같은 반구조화된 정보의 분석의 필요성 부각.

  웹문서의 시스템적 분석을 위해서 정보추출, 텍스트마이닝을 통한 분류, 키워드 추출, CO-word 분석 등의 적용이 시도되고 있음.

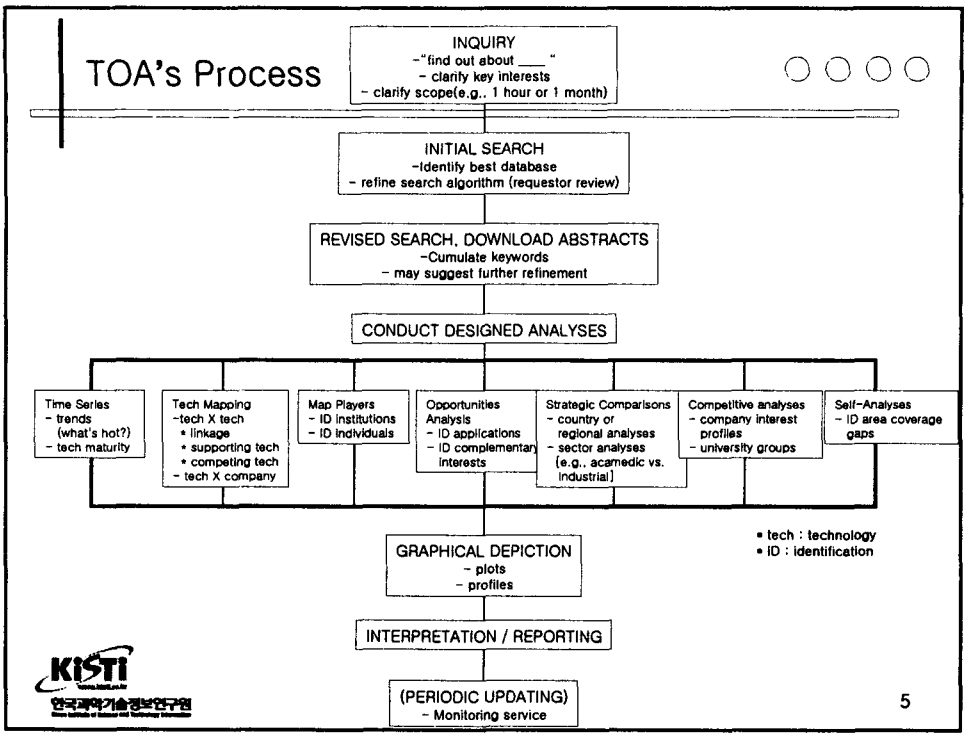
정보분석 시스템은 ?
특정 기술의 과거 발전 추세 분석, 기술예측, 특정 기술분야에 새로운 기술이 개발되었을 때 이를 자동으로 알려주는 것을 목표로 하며, 이를 위해서 과거의 추세와 앞으로의 변화를 도식화하고, Emerging technology의 핵심기술을 알려주고, 요소기술간의 상관관계(관련도)를 연산하고, 수많은 문헌 중에서 이러한 변화를 인지할 수 있는 정보를 담고 있는 문헌을 별도로 구분하여 요약과 함께 제시해주는 기능을 갖추어야 함.

정보분석 시스템화 사례조사 - 1

TOA (Technology Opportunities Analysis)

- TOA는 출판물, 특허, 인용, 프로젝트 데이터베이스의 초록에서 특별한 기술혁신의 전망에 관한 유용한 정보를 추출할 수 있다는 전제하에 Georgia Tech에서 1990년부터 개발하여 왔다. (Porter 등)

- TOA가 추구하는 목표 기능
  - 요소기술을 찾아내고, 그 요소기술들이 서로 어떻게 연관을 맺는지 밝혀낸다.
  - 그 기술을 누가(회사, 대학, 개인) 활발하게 개발하고 있는지 밝혀낸다.
  - 활발하게 개발하고 있는 사람들이 국내 또는 국제적으로 어디에 존재하고 있는지 밝혀낸다.
  - 시간이 경과함에 따라 기술적 강조가 어떻게 이동하고 있는지 밝혀낸다.
  - 연구 윤곽을 확인함으로써 조직의 강점과 약점을 밝혀낸다.



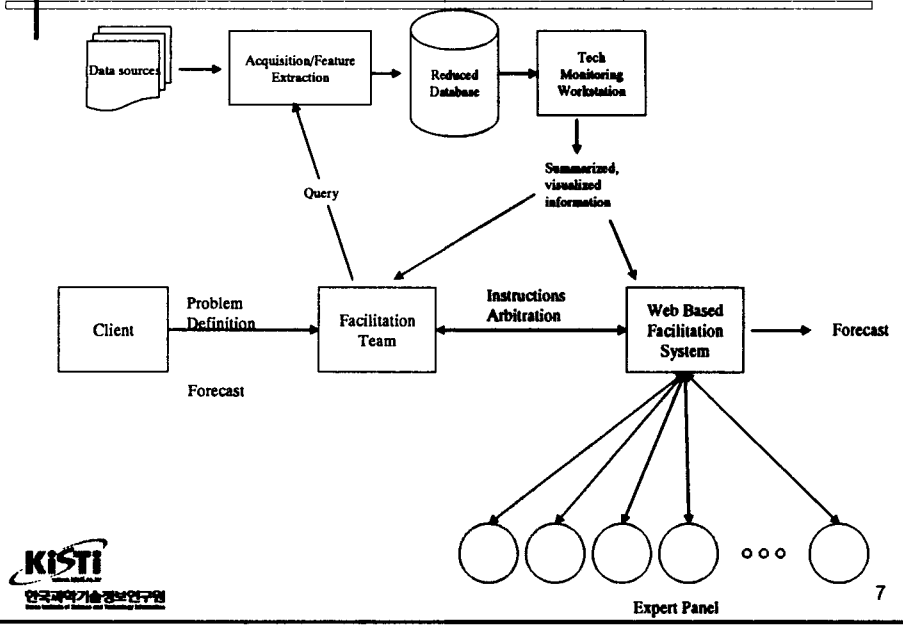
정보분석 시스템화 사례조사 - 2

DIVA (Database Information Visualization and Analysis system)

- DB의 문헌들 간, 문헌들의 클러스트들 간의 관계 변화(trend)를도식화해줌으로써 기술예측을 가능하게 해주는 프로그램
- 문헌간의 관련성의 크기(유사도, similarity)를 측정하는 문제를 해결하는 방법으로서 co-occurrence similarity function 제시
- 클러스터를 형성하는 방법으로서는 문헌간의 유사도에 의해 사각평면상의 상대적 위치(거리)를 결정하는 documents mapping 제시
- force directed placement와 self-organazing map(SOM) 기술을 토대로 documents mapping 처리
- Steven 등이 제시

KISTI
6

DIVA's Process

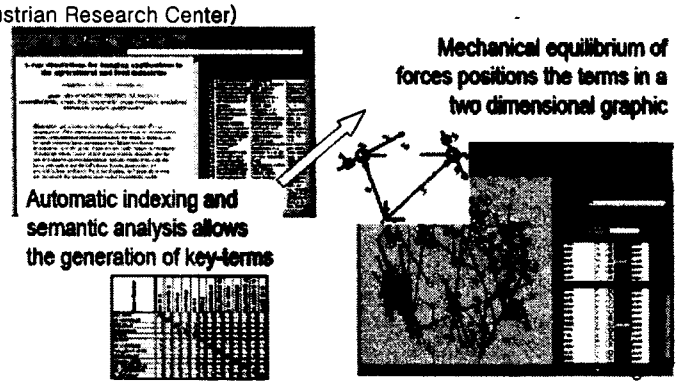


7

정보분석 시스템화 사례조사 - 3

BibTechMon

- 입력 받은 텍스트 정보를 자동 분류하고 키워드를 추출해서knowledge map을 자동 생성하는 기능
- 특히 특허 정보를 분석하는데 유용하지만, 기타 텍스트 정보를 내부 DB로 받아들이는 기능이 있기 때문에 특허 외의 문헌정보의 분석도 가능
- Margit 제안 (Austrian Research Center)



정보분석 시스템화 사례조사 - 결과종합 ○○○○

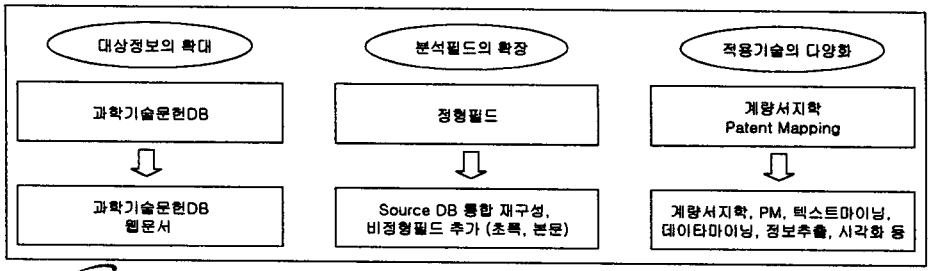
- 데이터수집(정보검색) → 분석대상 과제의 Data set 형성 → 데이터처리 → Visualization
- 데이터처리 과정에는 마이닝을 통한 키워드추출, 클러스터링, 문헌(특허)간 연관관계 분석 등을 거치고, Visualization은 분석 결과를 Map으로 도시화 하는 과정이 공통적으로 포함

시스템명	데이터수집	데이터처리	Visualization	분석결과정리
TOA	Search and retrieve DB	identify the term clusters→build a similarity matrix of the clusters	represent the similarity matrix in low-dimensional mapping	Tech. Forecasting
Tech OASIS	Search and retrieve text information	Profile the resulting search set→Extract latent relationships	Represent relationships graphically	Interpret the prospects for successful technological development
DIVA	Project DB	Similarity Function →Mapping→ Clustering→ Indicator	Visualization	Reports
Xavier Polanco	Data Selection	Term Extraction and Filtering →Data Clustering and Classification	Mapping or Visualization	Result Interpretation
BibTechMon	Database searches	Build-up of a database containing relevant information→Automatic keyword generation	Generation of knowledge maps based on co-word analysis	Cluster analysis
SOFM PM	collect patent document	Text Mining→Transform into structured data→ PCA→Reduce dimension Select relevant vector	SOFM(Self-Organizing Feature Map-based)	Develop PM
특허인용분석	특허문서의 수집	Text Mining→구조화된 자료로 변환→특허간 연관관계분석	네트워크 도시	특허 분석

정보분석시스템 모델 개발 ○○○○

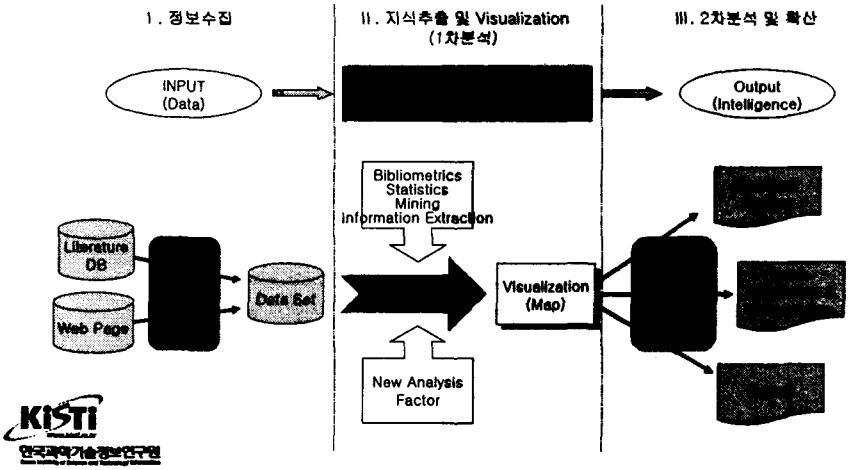
- 정보분석 시스템화는 전통적 정보분석을 대체하는 것이 아니라 각 기술분야의 전문가(또는 정보분석가)가 정보분석을 수행하는데 있어서 정보분석의 객관화, 계량화, 가시화 할 수 있는 정보를 창출함으로써 정보분석의 생산성을 높여주고, 한편에서는 정보분석 미숙련자도 정보분석결과를 쉽게 이용할 수 있도록 하는 것을 그 목적으로 함.

- 정보분석시스템 모델의 개발방향 : 가용한 모든 정보자원과 관련 기술 적용, 대상필드도 텍스트 마이닝 등을 적용하여 비정형필드 추가



정보분석시스템 모델 개발- 정보분석과정○○○○

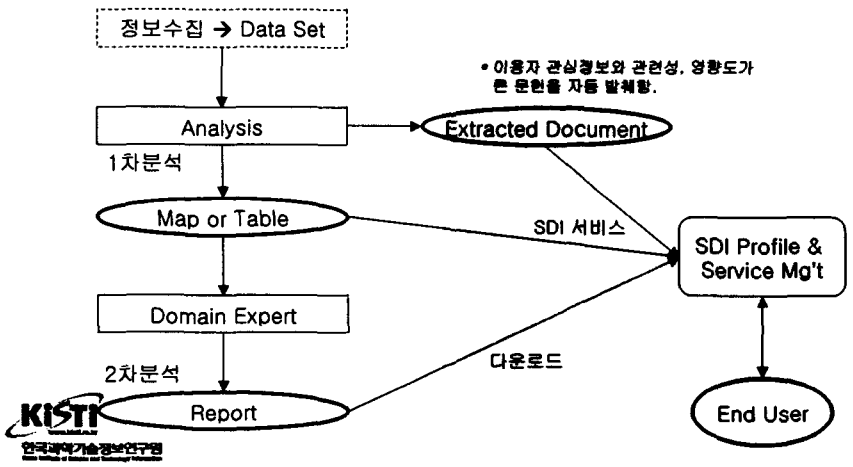
- 정보수집 및 2차분석 단계에서 Domain Expert의 개입을 필요로 하는데, 개입량을 최소로 하고 개인 편차의 영향을 줄임.
- 1)information professional, 2)individual contributor, 3)manager 관점 반영



KISTI 한국과학기술정보연구원

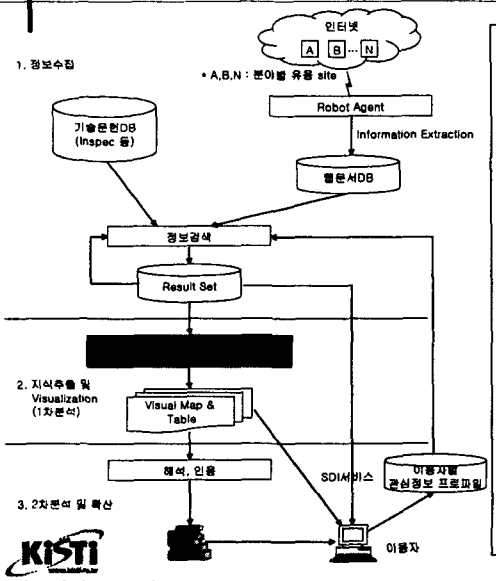
정보분석시스템 모델 개발- 2차분석 및 확산 과정○○○○

- 1차분석 정보, 2차분석 정보에 대한 SDI서비스 실시
- 특히 Extract Doc.와 MAP 정보를 정보분석 비 전문가도 활용할 수 있는 모델 구축



KISTI 한국과학기술정보연구원

정보분석시스템 모델 개발- "모델" ○○○○



- 웹문서를 분석대상에 포함한다.
- 키워드 리스트를 자동 생성한다.
- Domain expert의 개입을 최소화 할 수 있는 모델을 개발한다.
- 검색 Agent를 적용하여 서로 다른 포맷의 정보원을 쉽게 활용할 수 있도록 한다.
- 관심주제를 비중있게 다룬 문헌을 발췌하여 자동 서비스한다.
- Tech monitoring & forecasting. 요소기술간의 관계, 이슈기술의 출현, 조직의 기술적 강점과 약점을 규명한다.
- 이용자 그룹을 information professional individual contributor, manager로 구분하여 서비스 충족 방안을 다차원적으로 모색한다.
- 로봇에이전트, information extraction, 정보검색 에이전트, 텍스트마이닝, SDI서비스 등 관련 기술을 적용한다.
- 구성요소들을 장기적 목표를 세우고 개발한다.

KISTI 한국과학기술정보연구원

13

구성기술 고찰 - 정보수집용 로봇에이전트 ○○○○

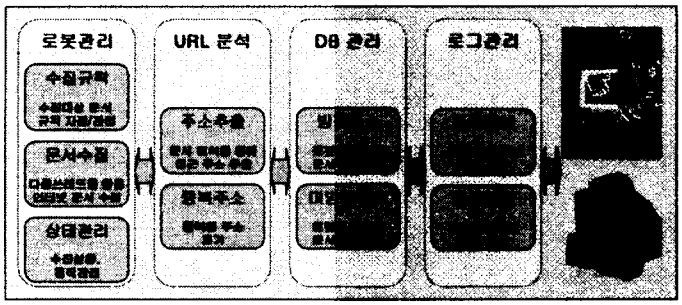
- 웹 로봇은 인터넷상의 정보수집을 위해 사용되는 하나의 에이전트로서 서버에 상주하면서 사용자와 직접적인 상호작용 없이 사용자를 대신해서 작업을 수행하여 인터넷상에서 분산된 온라인 정보를 순회하며 정보를 수집하는 프로그램이다.
- 웹 로봇이 가장 많이 사용되는 부분은 웹 상의 대규모 데이터를 대상으로 문서를 수집하고 수집된 문서에서 필요한 자원을 발견하는 데 있다고 할 수 있다.
- 최근에는 전문정보를 대상으로 하는 전문 포탈사이트가 구축되면서 일반적인 웹 로봇에 전문정보를 선택하기 위한 다양한 언어처리 기술이 도입되어 사용되고 있다.
- 또한 로봇은 다양한 기능을 동시에 수행하기도 한다. 어떤 로봇 에이전트는 특정 홈페이지의 내용을 차례차례 방문하고, 그 내용을 분석하여 알려줌과 동시에 특정 페이지 자료만을 가지고 키워드에 의한 검색이 가능하게 하기도 한다. 또한 특정 문서 정보만을 추출하기 위한 필터와 연결되어 문서 수집시 관련된 페이지만 접근하여 수집대상 페이지에 대해서 수집하기 위해 사용된 경우는 이에 대한 별도의 DB 관리 프로그램과 연결되어 자동 키워드 및 정보추출 기능을 동시에 행하고 있다.

로봇의 특성 중에서 본 연구 관련 기능과 동향만 나열

KISTI 한국과학기술정보연구원

14

구성기술 고찰 – 로봇에이전트의 구성요소 ○○○○



- 로봇관리 : 사용자 인터페이스로부터 사용자의 입력을 통해 로봇 에이전트를 수행하고 관리하는 역할.
- URL 분석 : 로봇이 접근해야 될 주소정보를 추출하기 위해 사용된다.
- DB 관리 : 로봇이 방문한 주소와 URL 분석에 의해 추출된 미방문 주소를 별도로 관리하고 미방문 주소를 로봇관리 모듈의 문서수집기에 전달하는 기능을 수행한다.
- 로그관리: 로봇의 동작과 관계된 정보를 로그정보를 기록하는데 사용된다. 저장된 로그분석을 통해 로봇의 동작에 대한 통계정보를 생성할 수 있다.

구성기술 고찰 – 로봇에이전트의 문제점/해결방안 ○○

- 웹 상의 문서는 사용자가 보기에 편하도록 설계되어 생성된 것이지 프로그램의 해독성을 편리하도록 만들어진 것이 아니기 때문에 필요한 정보만을 뽑아내는 정보추출은 매우 어려운 작업이다.
- 특정 사이트에서의 웹 문서의 구조가 다른 사이트에서도 똑같이 적용되는 경우는 매우 드물다. 따라서, 적용 사이트가 하나 추가될 때마다 정보추출 모듈을 하나 더 만들어 주어야 하는 부담이 생긴다.
- 정보추출 대상이 되었던 사이트들은 정적으로 존재하는 것이 아니라 웹 페이지의 특성상 동적으로 자주 페이지 구조를 변경하기 때문에 이때마다 웹 페이지 설정사항을 갱신해야 하는 어려움이 있다.
- 다국어 지원을 하지 못하여 영어, 한국어 이외의 정보는 수집이 불가능하다.
- 수집하고자 하는 정보가 검색결과일 때 검색 후 수집이 불가능하다.

해결과제

- ●유동적인 웹 문서로부터 정보 추출의 무결성을 보장하기 위해 기존의 웹 로봇 보다는 강력하고 유연성을 가진 웹 로봇 개발이 필요하다.
- ●수집되는 웹 문서를 특정 키워드에 의해 분류하는 규칙기반 분류방법 보다는 내용분석과 동시에 글의 문체나 스타일에 의해 분류할 수 다차원 분류기술이 필요하다.
- ●수집 대상의 웹 문서에 의미정보를 포함하여 수집할 수 있도록 하는 에이전트 기술의 접목이 필요하다.

구성기술 고찰 : 텍스트마이닝(1) - 정의 ○○○○

- 데이터마이닝이 구조적인 데이터를 대상으로 유용하고 잠재적인 패턴을 끌어내는 것이라고 한다면, 텍스트마이닝은 자연어로 구성된 비구조적인 텍스트 안에서 패턴 또는 관계를 추출하여 지식을 발견하는 것으로, 주로 텍스트의 자동 분류작업이나 새로운 지식을 생성하는 작업에 활용되고 있다.

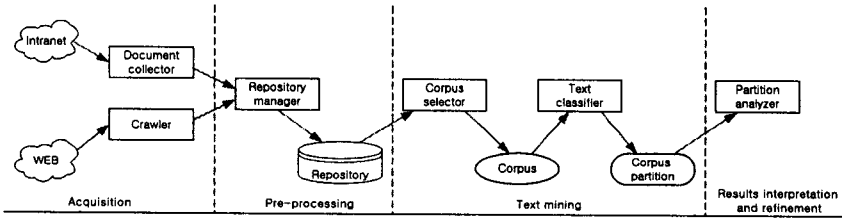
- 오늘날 정보를 저장하는 가장 자연스러운 기본적인 형태는 결국 "텍스트"이고, 최근 조사에 따르면 기업체 정보의 80%가 텍스트 문서 형태로 보관되고 있기 때문에 자연어로 된 텍스트문서의 자동화되고 지능적인 분석은 매우 중요하다.

- 텍스트마이닝은 데이터 준비 단계(Text Refining)와 지식 추출(Knowledge Distillation) 과정으로 나타낼 수 있다.

  - 데이터 준비단계 : 다양한 정보원(Information source : 인터넷, 인트라넷, 이메일 등)에서 수집한 자유로운 형태의 텍스트 문서를 Intermediate Form으로 바꾸는 단계로서 자연어처리, 웹문서의 경우 태그 제거, URL과 타이틀 추출 등의 기술체계를 포함하고 있다.
  - 지식추출과정 : Intermediate Form의 문서에서 의미있는 패턴과 지식을 유추해 내는 과정으로서 클러스터링, 분류(Classification), 시각화(Visualization), 문서요약, 기계학습 등의 기술체계를 포함하고 있다.

KiSTi 한국과학기술정보연구원

구성기술 고찰 : 텍스트마이닝(2) - 진행과정 ○○○



1. 문서수집(Document acquisition)

- 텍스트마이닝은 기업(기관) 내외부에서 발생하는 모든 텍스트 문서를 수집 대상으로 한다. 웹에서 특정 기술분야(또는 학술분야)의 정보를 수집할 때에는 적절한 사이트로 수집범위를 제한하지 않으면 안된다.
- 적절한 사이트를 선정하기 위해서는 ①웹사이트 평가표에 의해 사람이 직접 평가하는 방법과 ②인터넷 검색엔진에서 적용하고 있는 사이트 랭킹 기술을 이용하는 방법 ③두가지 방법을 적절히 절충하는 방법에 대한 연구가 요구된다. 웹문서 수집을 위해서 Web Crawler 또는 로봇에이전트를 사용하는 방법은 이미 보편화되어 있다.

KiSTi 한국과학기술정보연구원

구성기술 고찰 : 텍스트마이닝(3) - 진행과정

2. 문서 전처리(Document pre-processing)

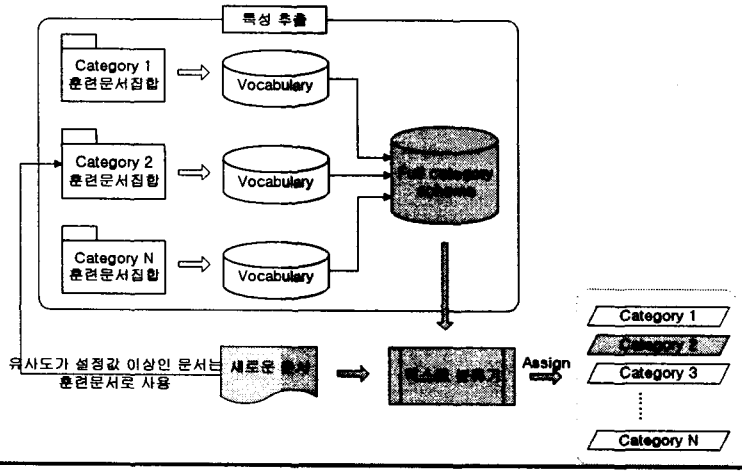
- 이 단계에서는 각 문서를 텍스트마이닝에 적합한 형태로 변환한다. 텍스트 자체가 자연어로 되어있기 때문에 언어학적 관점에서의 자연어 처리과정은 필수적이다.
- 이러한 전처리 과정을 통해 분석과정에 적합한 최적의 데이터 상태를 만들어 분석의 질을 향상시킬 수 있는데, 전처리 작업은 실제 분석에 소요되는 시간보다 더 많이 걸릴 수 있으며 수집한 데이터를 잘 이해하는 일이 필요하다.

3. 텍스트마이닝

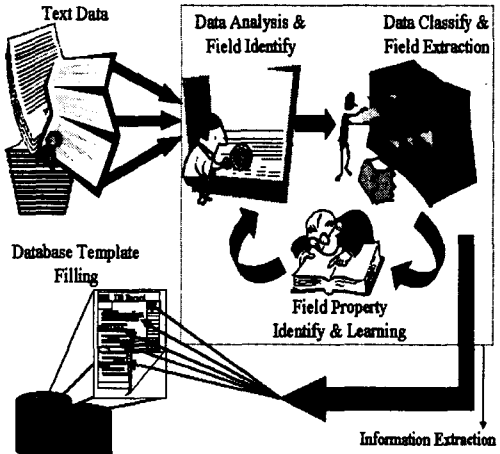
- 텍스트마이닝에 적용하는 주요 분석기법은 군집(Clustering)과 분류(Categorization)이며, 분류분석 수행에 앞서 군집화를 먼저 수행시켜 전체 문서집합의 개요를 획득하고 분류를 위한 판단기준을 얻어낸다.
- 군집은 분류의 준비단계로서 사용자가 i)분류해 낼 항목(Category)을 명확히 정의하고 ii) 각 항목에 따른 훈련문서를 선정하여 학습시키는 과정에 군집결과를 이용하는 것이다.

구성기술 고찰 : 텍스트마이닝(4) - 응용:텍스트분류

- 텍스트 분류(Text categorization)란 텍스트의 내용에 따라 미리 정의해놓은 범주를 부여하는 과정이다. 분류를 수행하기 위해서는 각 항목을 위한 학습데이터를 사용자가 선정하여 훈련시키는 과정이 필요하다. 선별된 각 훈련문서에서 특성을 추출해 내어 특성벡터(feature schema)를 구성하는 기본 학습과정을 포함.



구성기술 고찰 : 정보추출(1)



- 정보추출 : 한 문서에서 그 문서의 중심적 의미를 나타내는 특정 구성요소를 인식하여 추출하는 작업
- 추출된 정보는 필드 제한검색, 마이닝 적용 등에 의해 가용성을 높이기 위하여 데이터베이스에 필드별로 저장
- 인터넷의 정보추출 작업의 해결과제
  - ① 정보소스의 문서들은 원칙적으로 사람들이 읽기 편하도록 작성되었기 때문에 프로그램이 쉽게 처리할 수 있도록 문서 구성 포맷 관행에 대한 정보를 제공해주는 사이트는 거의 없다. → XML (?)
  - ② 한 사이트에서 사용된 독특한 포맷 관행이 다른 사이트에도 적용될 가능성이 거의 없기 때문에 사이트가 추가되는 경우 새로운 wrapper (추출규칙)를 구성해야 한다.
  - ③ 사이트들이 자주 포맷을 바꾸고 있으며, 이전에 만들었던 wrapper가 동작하지 않게 된다. 이러한 변화를 즉시 감지하여,

KISTI 한국과학기술정보연구원

구성기술 고찰 : 정보추출(2)

- 정보추출의 관점에서 텍스트 문서는 구조화되지 않은(unstructured) 문서, 준구조화된(semi-structured) 문서, 구조화된(structured) 문서의 형태로 구분할 수 있다.
  - ① 구조화되지 않은 문서는 어떤 일정한 형식 없이 정보를 표현하는 방식으로 일반 텍스트의 경우 구조화 되지 않은 문서라 할 수 있다.
  - ② 준구조화된 문서는 일부 정보는 구조화 되어있고 다른 일부는 비구조화 문서로 이루어지 문서를 말한다.
  - ③ 구조화된 문서는 정보를 테이블 형태와 같이 일정한 구조로 표현한 문서이다.
- 일반적인 웹상의 문서들은 구조화되지 않은 문서들이지만, 특정 도메인의 정보검색을 위한 문서들은 준구조화 문서로 볼 수 있다.

[정보추출적용례 : 전자상거래 Site]

Input Data

> Recently, LG Electronics announced Platon flat TV. Its model name is CN-230, and the price of 29" is 1000$. This TV has full-flat dark tint tube.

⇩

IE System

⇩

Output Template

item	: plat flat TV
manufacturer	: LG Electronics
model	: CN-230T
price	: 1000$
size	: 29"
specification	: full-flat dark tint

KISTI 한국과학기술정보연구원

22

결론 및 향후 연구 – 진행중 연구과제 ○○○○

- 정보분석 자원 발굴 프로세스에 관한 연구
  - 과학기술문헌 정보를 제공하는 우량 사이트를 발굴하는 모델 개발
  - 본 연구의 가정 : 분야별로 좋은 사이트 몇군데만 조회해도 그렇지 않은 대부분의 사이트 전체를 조회하는 것 보다 우량 정보를 획득할 확률이 높다.
  - 발굴 프로세스의 확장성 검증 : 특정 기술분야 또는 산업분야(예:생물산업)를 대상으로 실증적 연구 진행

- 정보분석 인자 추출에 관한 연구
  - 궁극적으로는 과학기술문헌DB와 웹사이트를 검색결과를 온라인상에서 시스템적으로 신속하게 분석함으로써 연구개발(기술개발)의 동향을 파악하거나 발전방향을 예측할 수 있는 시스템을 구현하고자 한다.
  - 본 연구에서는 이러한 시스템을 구현하는데 기초가 되는 정보분석인자를 조사하고, 새로운 분석인자를 발굴하는 것을 목적으로 한다.
  - 특히 KISTI에서 진행된 과거의 관련 연구결과를 데이터마이닝(또는 텍스트마이닝) 관점에서 문제점/개선점을 고찰하고, 텍스트마이닝 등을 적용하여 비정형필드에서의 인자추출이 가능한지에 대한 연구를 하고자 함.
  - 새로운 인자의 확장성 검증 : INSPEC에 대한 인자를 추출하는 실증적 연구 진행

KISTI 한국과학기술정보연구원

결론 및 향후 연구 ○○○○

- 텍스트마이닝의 기술적 연구 과제
  - 텍스트마이닝이 자연어 처리에 기반을 두고 있기 때문에 의미 파악을 위한 계산이 복잡하여 처리시간이 과다하게 소요되고, 분석 후 오류율도 많이 개선할 여지를 가지고 있다.
  - 2개 이상의 언어를 동시에 인식하고 궁극적으로는 다국어를 동시에 처리할 수 있도록 성능이 개선되어야 한다.
  - 각 웹사이트의 규칙성을 감안하여 파싱을 할 경우 비구조적 텍스트 문서를 구조화 하는게 가능하다. 텍스트마이닝 시스템이 이와 같이 domain knowledge를 어떻게 잘 이용해서 parsing efficiency를 개선할 수 있을지, 그리고 보다 더 compact 한 intermediate form 을 만들어 낼 수 있을지에 대한 연구가 필요하다.
  - 현재까지는 전문가들을 대상으로 한 텍스트마이닝 툴이 있을 뿐이지만, 향후에는 비전문가들도 쉽게 사용할 수 있는 툴이 개발되어야 한다.

- 웹문서 정보분석시스템의 한계 해결
  - 웹문서는 저널과 같이 발생년도(출판년도)가 명확하지 않기 때문에 원저작이 이루어진 년도를 가지고 시계열적 관점에서 정보분석을 하는 것이 어렵다.KITAS의 정보분석의 경우 14개의 분석인자 모두가 연도를 인수로 한다.
  - 정보추출에 의해 비구조화된 문서를 구조화 하여 필드별로 구분하여 DB를 생성하기위해서는 각 도메인에 대한 지식을 필요로 하고, HTML 문서의 태그 구분 부족때문에 정보추출에 한계가 있다. PDF 또는 워드파일의 경우 텍스트로 변환하게 되면 정형필드로 지정해야하는 제목, 저자, 연도 등의 폰트 사이즈라던가 공간상에서의 위치 정보가 사라지기 때문에 형식 식별이 어렵다.
  - 웹사이트의 경우 도메인에 대한 지식을 확보하여 이를 바탕으로 정보추출(Information Extraction)를 하는 데에는 현실적 한계가 있고, 웹사이트가 XML과 같은 구조화된 언어로 변환되지 않는 한 정형 필드를 대상으로 실시하는 정량적 분석은 적용하기 어렵다.
  - 이러한 한계 때문에 웹정보 분석시스템의 모델은 문헌 DB 분석시스템에서 채택한 모델의 일부만을 변형해서 적용할 수 밖에 없고, 특히 문서의 발행연도를 인수로 하는 분석은 거의 불가능하다.

KISTI 한국과학기술정보연구원

참고문헌

- Robert J. Watts et al. "Factor Analysis Optimization: Applied on Natural Language Knowledge Discovery ", http://www.tpac.gatech.edu/papers/FA_Opt2.pdf, pp. 1-13.
- *Mrrgit Noll etal, "Knowledge maps of knowledge management tools - Information visualization with* BibTechMon", PAKM 2002(Vienna, 203, Dec. 2002)
- Yoon, Byung-Un, Yoon, Chang-Byung, Park, Yong-Tae(2002), "On the development and application of a self-organizing feature map-based patent map", R&D Management, 32~4, pp. 291-300.
- 윤병운 외 (2001), 「데이터마이닝을 이용한 특허 인용 분석」, 한국경영과학회/대한산업공학회 춘계공동학술대회(2001년 4월 27일) 발표 논문지, pp.583-586.
- Leon M. Galitsky et al.(2003), "A survey of emerging trend detection in textual data mining", Survey of text mining, pp.185-224.
- "Text Mining, Review of TPAC Technologies for ONR", 2002.8, pp.1-3.
- Morris, Steven et al.(2002), "DIVA: A Visualization System for Exploring Document Databases for Technology Forecasting", Computers & Industrial Engineering, Sep., 43-4, 841-862.
- Yoon, Byung-Un, Yoon, Chang-Byung, Park, Yong-Tae(2002), "On the development and application of a self-organizing feature map-based patent map", R&D Management, 32~4, 291-300.
- 문영호 외(2000), 「온라인 DB검색을 통한 기술분석시스템 구축」, 산업자원부.
- Ah-Hwee Tan (1999), "Text Mining : The state of the art and the challenges", http://textmining.krdl.org.sg/people/ahhwee/.
- IBM Systems Journal, vol.40, no.4, 2001, p.968-969
- 최윤정 외(2002), 「웹 컨텐츠의 분류를 위한 텍스트마이닝과 데이터마이닝의 통합 방안 연구」, 한국인지과학회 논문지, 제13권, 제3호, pp.33-46.
- 조태호(2001), 「텍스트마이닝의 개념과 응용」, 지식정보인프라, pp.76-85.
- Berry de Bruijin(2002), Int. Journal of Medical Informatics, no.67, pp.7-18.
- Dorre J., P. Gerstl, R. Seiffert(1999), "Text Mining: Finding Nuggets in Mountains of textual Data", in Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining
- 김재훈, 장병탁, "온닉 마코프 모델을 이용한 정보 추출", 서울대학교 컴퓨터공학부

KiSTi 한국과학기술정보연구원