국가R&D정보를 활용한 기업 대표 키워드 DB 구축 방법

Enterprise Representative Keyword Database Construction from National R&D Information Collection

한 희 준, 김 병 정, 최 희 석, 김 재 수
한국과학기술정보연구원 NTIS센터

Heejun Han, Byeongjeong Kim,
Heeseok Choi, Jaesoo Kim
NTIS Center, Korea Institute of Science and Technology Information

요약

기업이 원하는 R&D정보를 추출하기 위해서는 R&D정보 검색에 활용할 질의어가 있어야 한다. 먼저 구축되어야 한다. 기업마다 관심있는 제품과 기술 키워드가 각각 다르다. 기업에 적합한 R&D정보를 생성하기 위해 질어어로 사용될 기업을 대표하는 키워드 군을 생성하고자 한다. 본 논문에서는 2002년부터 기업이 수행한 국가 R&D과제정보와 과제에서 도출된 논문, 특허, 연구보고서 등 성과정보로 부터 기업을 대표하는 키워드를 추출하고 이를 웹에서 크롤링한 기업정보와 비교하여 기업 대표 키워드 데이터베이스를 구축하는 방안에 대해 논한다.

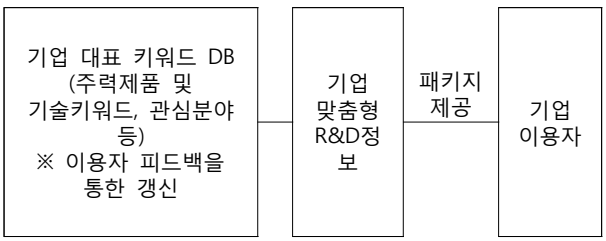
I. 서론

기업은 국가연구개발 업무를 수행하는데 있어, 사전조사, 기획, 과제신청, 연구개발수행, 사업화 및 기술이전이라는 일련의 단계를 가진다. 현재 여러 부처에서는 중소기업 성장 활성화를 위해 기업을 지원하는 여러 서비스를 운영중이다. 특히 국가과학기술지식정보서비스(이하 NTIS)는 17개 부처·청으로부터 국가R&D와 관련된 종합정보를 수집하여 서비스함으로써 기업의 연구개발활동을 지원하고 있다[1]. 하지만 이용자는 과제정보, 참여인력정보, 논문, 특허, 연구보고서, 사업화, 기술이전 등 성과정보, 기술산업 정보 등 원하는 국가R&D정보를 이용하기 위해 NTIS 서비스에 직접 접근하여 관심영역에 대한 검색을 수행해야만 한다. 기업 이용자는 이용하는 서비스에 접근해 원하는 정보를 찾는데 많은 어려움을 가지고 있으며, 본인이 원하는 정보를 서비스가 스스로 작성하여 제공받길 원한다[2].

기업은 대부분 주력제품 및 기술을 보유하고 있고, 이를 활용해 이익실현을 추구한다. 기업마다 해당 기술 및 제품과 관련된 특정 키워드가 존재한다. 본 논문에서는 기업이용자가 원하는 정보를 생성하기 위해 기업 대표 키워드를 구축하는 방안에 대해 논한다. 먼저 NTIS 보유정보로부터 기업 대표 키워드 군을 추출한 후 웹크롤링 기법을 이용해 수집된 기업의 제품 및 기술 키워드와 비교한 후, 마지막으로 이용자 피드백을 통해 키워드를 갱신하는 방안을 논한다.

Ⅱ. 기업 대표 키워드 추출 방법

NTIS는 기업이용자가 검색을 수행하지 않고도 미리 기업이 원하는 정보를 만들어 제공하고자 한다. 그림 1은 기업 대표 키워드를 추출하여 이를 활용해 기업 맞춤형 R&D정보를 생성하고 이를 제공하는 NTIS 기업 R&D정보제공 개념도이다. 본 장에서는 기업 맞춤형 정보를 생성하는데 활용할 기업 대표 키워드 추출방법에 대해 제안한다.



▶▶ 그림 1. 기업 맞춤형 R&D정보 제공 개념도

1. 국가 R&D 정보에서 기업 대표 키워드 추출

NTIS는 정부부처, 국공립연구소, 출연연구소, 대학 및 기업에서 수행한 국가 R&D과제와 과제로부터 도출된 논문, 특허, 연구보고서, 사업화, 기술이전 정보 등의 성과정보를 보유하고 있다. 이 가운데 2002년부터 대기업 및 중소기업이 연구주관기관 또는 협동연구기관으로 참여한 국가R&D과제의 수는 116,881건으로 전체 과제 대비 20%에 달하는 수준이다.

대기업과 중소기업이 수행주체가 된 과제정보로부터 기업 대표 키워드를 추출하기 위해, 과제의 메타정보 가운데 과제명, 과제내용, 목표, 기대효과, 과제키워드를 활용한다. 특정 기업 A가 참여한 과제로부터 해당 메타데이터를 공기(white space) 기반으로 나열한 후, NTIS 가

보유한 과학기술전문용어사전을 통해 기술용어를 먼저 식별한다. 이 때 조사, 동사, 형용사 등의 불필요한 단어는 제거되며 기술용어로 활용 가능한 단어를 생성한다. 이 단계에서 A 기업이 수행한 과제 메타정보에서 추출한 기술용어는 수백 단어 이상이 될 수 있다. 그 후 단어의 빈도(frequency) 수를 계산하여 자주 출현한 기술용어 순으로 랭킹을 계산한 후 상위 10개의 후보군을 생성한다. 예를 들어 (주)서울반도체는 그 간 35건의 국가R&D 과제에 대해 주관기관이나 협동연구기간의 역할을 수행하였으며, 해당 과제로부터 추출한 기술용어 상위 10개는 아래와 같다.

- LED, 광결정, PPFD, 나노기반발광다이오드, 공명에너지전달, 나노로드, 나노패터닝, Acrich, 조명, 질화갈륨

2. 기업 홈페이지 크롤링을 통한 기업 대표 키워드 보완

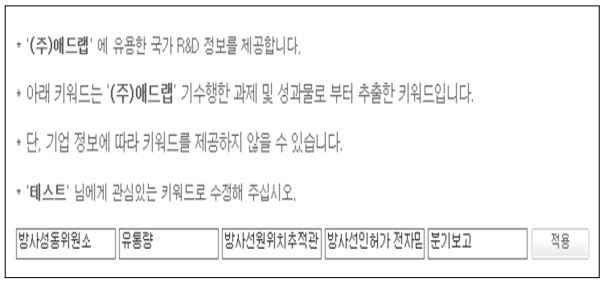
대부분의 기업은 당사를 홍보하고 제품 및 기술을 제공하기 위해 홈페이지 등의 웹서비스를 운영한다. 주력제품 및 보유기술에 대한 정확하고 최근 정보를 웹을 통해 제공하는데, 이는 최근 기업의 관심 키워드를 추출하는데 있어 유용한 정보이다. 웹크롤링 기법을 활용하여 해당 기업의 홈페이지 소스를 수집하고, 수집된 html 기반의 텍스트를 NTIS 과학기술전문용어사전으로 필터링하여 키워드 리스트를 추출한다. 역시 추출된 키워드 리스트 가운데 빈도 수를 계산하여 상위 5개 단어를 활용한다. 기업의 최근 주력제품 및 기술 영역으로 판단 가능한 키워드 리스트는 기업 대표 키워드로 활용하기에 적합한 단어들이다. (주)서울반도체의 경우 기업 홈페이지 크롤링과 대표 키워드 추출 결과 'LED', 'Acrich', '발광다이오드', 'nPola', '발광', 'Power' 가 추출되었다. 그림 2는 (주)서울반도체 홈페이지의 제품정보 페이지를 크롤링한 예시이다.

```
<div class="cate1"><!-- cate1 -->
  <h3 class="arrowH3"><a href="/kr/html/Product/product_view.asp?catecode=1013">Acrich Series</a></h3>
  <ul>
    <li><a href="/kr/html/Product/product_view.asp?catecode=1013">- Acrich MJT</a></li>
    <li><a href="/kr/html/Product/product_view.asp?catecode=1002043">- Acrich2</a></li>
    <li><a href="/kr/html/product/Product.asp?catecode=1003">- Acrich</a></li>
  </ul>
  <h3 class="arrowH3 mt5"><a href="/kr/html/product/product_view.asp?catecode=2004005">LEDs by Type</h3>
  <ul>
    <li><a href="/kr/html/product/product_view.asp?catecode=2004031">- Z-Power LED</a></li>
    <li><a href="/kr/html/product/product_view.asp?catecode=2005038">- Top View LED</a></li>
    <li><a href="/kr/html/product/product_view.asp?catecode=2006024">- Through Hole</a></li>
    <li><a href="/kr/html/product/Product.asp?catecode=2007">- Side View LED</a></li>
    <li><a href="/kr/html/product/product_view.asp?catecode=2008026">- Chip/Sensor</a></li>
    <li><a href="/kr/html/product/module.asp">- Customized Module</a></li>
    <li><a href="/kr/html/product/product_view.asp?catecode=2012">- Chip On Board</a></li>
  </ul>
  <h3 class="arrowH3 mt5"><a href="/kr/html/product/power.asp?gubun=3">LEDs by Power</a></h3>
  <ul>
    <li><a href="/kr/html/product/power.asp?gubun=3">- High Power</a></li>
    <li><a href="/kr/html/product/power.asp?gubun=2">- Mid Power</a></li>
    <li><a href="/kr/html/product/power.asp?gubun=1">- Low Power</a></li>
  </ul>
  <h3 class="arrowH3 mt5"><a href="/kr/html/Product/npola.asp?catecode=301">nPola</h3>
  <ul>
    <li><a href="/kr/html/Product/npola.asp?catecode=301">- nPola</a></li>
  </ul>
```

▶▶ 그림 2. 기업 웹페이지 크롤링 소스 예시

3. 기업이용자 피드백을 통한 기업 대표 키워드 구축

위 장에서 추출된 키워드 15개 가운데 중복은 제거하고, 중복 키워드가 존재하지 않을 경우 2.2에서 추출한 5개 대표 키워드와 2.1에서 추출한 상위 5개 키워드, 10개의 기업 대표 키워드가 이용자에게 제시된다. NTIS 는 이용자가 로그인 할 때 기업 이용자의 경우 회원정보로부터 해당 소속기관명을 식별하고, 해당 기업명에 매칭되는 기업 대표 키워드를 제시한다. 이용자는 제시된 단어를 수정함으로써 본인의 관심연구분야 및 기술에 적합한 키워드를 관리한다. 그림 3은 위에서 설명한 추출과정을 통해 만들어진 기업 대표 키워드를 이용자에게 제시하여 갱신을 요청하는 화면이다.



▶▶ 그림 3. 기업이용자 피드백을 통한 기업 대표 키워드 구축

Ⅲ. 결론

NTIS는 17개 부처 · 청으로부터 R&D 정보를 수집하여 서비스하고 있으나, 기업 이용자는 본인의 관심분야를 미리 파악하여 서비스가 미리 관심정보를 제공해주길 원한다. 본 논문에서는 국가R&D정보를 패키징하기 위해 질의어로 활용할 기업 대표 키워드를 구축하는 방안에 대해 논하였다. 구축된 키워드들은 특정 기업이 원하는 국가 R&D 정보를 만들기 위해 질의어로 활용되는데, 검색을 위한 알고리즘 및 패키징 방안에 대한 연구가 추가적으로 필요하다. 또한 기업은 일정 주기별로 변하는 시장환경에 맞추어 기술개발 영역을 확대하거나 구체화 시키므로, 향후 연구에서는 기업 대표 키워드를 보완하기 위해서는 입수되는 과제나 성과정보로부터 주기적으로 키워드를 추출하거나, 이용자 피드백을 효과적으로 반영하는 방안이 필요하다.

▌Acknowledgement▐

본 연구는 '14년도 한국과학기술정보연구원에서 수행하는 '국가R&D 성과물 공동활용체제 구축' 사업의 지원을 받아 수행된 연구임. (K-14-L02-C02-S02)

▌참 고 문 헌▐

[1] 정연덕, "R&D 활성화를 위한 국가과학기술종합정보서비스(NTIS:National Technology Information System) 활용 방안", IT와 법연구, 제4호, pp.1-25, 2010.
[2] 서상혁, 이선영, 이병희, "국가 R&D정보 이용자의 고객가치 및 고객만족도 영향요인 분석", 기술혁신학회지, 제15권, 제4호, pp.837-861, 2012.
[3] 국가과학기술지식정보서비스(www.ntis.go.kr)