# [한글날기념 취재 지원자료: 한글 정보화와 검색 솔루션]

## 한국 지형에 강하다

IT 강국이라는 호칭이 무색하게도 대한민국에는 아직 세계적이라는 수식어를 댈 수 있는 소프트웨어 기업이 많지 않습니다. 소프트웨어 단일 품목으로 연 매출 100억원을 넘기는 것이 이례적으로 받아들여질 만큼 규모 면에서 영세함을 면치 못하고 있는 것이 현실입니다.
이런 까닭으로 기업용 제품에서 개인용 패키지 소프트웨어 시장에 이르기까지, 대부분의 분야에서 다국적 소프트웨어 기업들이 국내 시장을 쥐락펴락 하고 있습니다. 몇몇 글로벌 벤더의 소프트웨어 유지보수 비율 인상으로 한창 떠들썩했던 것도 이와 무관치 않아 보입니다.

그런데 외산 제품의 홍수 속에서도 흔치 않은 사례로 국내 소프트웨어 기업들이 주도권을 잡고 있는 시장이 있습니다. 바로 검색 관련 분야입니다.

전세계 웹 검색 시장을 석권한 구글과 야후를 제치고 엠파스, 다음, 네이버가 국내 웹 검색을 장악한 것처럼, 국내 기업용 검색 소프트웨어 시장에서도 코난테크놀로지, 다이퀘스트와 같은 국내 기업들이 오토노미, 패스트서치와 같은 글로벌 기업을 누르고 주도권을 쥐고 있습니다.

이는 국내 검색 솔루션 기업들이 국내 기업환경에 최적화된 제품을 개발하고 마케팅 활동을 펼쳐 얻어낸 결과물이기도 하지만, 검색 결과 정확도에 지대한 영향을 미치는 한국어 검색에 최적화되어 있는 기술력을 확보하고 있기 때문이기도 합니다.

STS2008(Search Technology Summit 2008)이라는 기업용 검색 솔루션 전문 기업 6개사가 성공적으로 개최한 행사에서도 이러한 경향이 잘 나타나는데, 각 기업들은 검색 품질에 지대한 영향을 끼치는 '국어의 정보화'를 위하여 학계, 연구기관과 활발하게 교류하고 있습니다.

## 한글을 잘 다뤄야 좋은 검색 솔루션

자연어는 컴퓨터 프로그램 언어와는 달리 자연적으로 발생한 언어를 말합니다.
이러한 자연어의 정보 검색을 위하여 사용자가 찾는 정보가 무엇인지 정확하게 파악하는 것이 필수적입니다.

국어는 1byte로 구성되는 영어와 달리 음절 단위로 단어를 구성합니다. 즉, 중국, 일본 같은 동아시아권과 마찬가지로 2byte의 언어입니다.
외국계 솔루션 벤더들이 제품의 한국 현지화를 위하여 반드시 거쳐야 할 과정이 2byte 입력입니다. A, b, c와 같은 1byte 를 입력하는 영어와 달리, 국어는 가, 나, 다 와 같은 음절 단위의 언어이기 때문에 2byte 입력 문제를 해결해야 제품의 제대로 된 현지화를 이룰 수 있습니다.

추가적으로 국어는 문장의 시제 역시 음절의 추가로 해결하는 절약의 언어입니다. 예를 들어 '만들다' 라는 동사의 '었' 혹은 '겠' 이 들어감으로써 시제가 바뀌게 됩니다. 이 때문에 정확한 검색결과를 보여주기 위하여 한국어 형태소 분석이 중요합니다.

형태소 분석은 색인 대상 또는 질의(쿼리)로부터 의미 있는 정보를 추출하기 위하여, 문장 또는 어절을 의미를 갖는 최소 단위인 형태소로 분리하는 과정을 말합니다.

*예를 들어 '감기는' 이라는 단어를 검색했을 경우, 기침을 동반한 '감기'인지, 비누로 머리를 '감기' 인지 알아야 정확한 검색 결과를 제공할 수 있습니다.*
이러한 형태소 분석은 분석 후보의 생성과 그 후보들로부터 옳은 분석 결과를 선택하는 과정으로 보유 단어가 많을수록 정확한 검색 결과를 산출할 수 있게 됩니다.
그래서 검색 기업들은 수백만 단어를 담은 정보 처리용 국어 사전을 구축하여 더 나은 결과를 제공할 수 있도록 노력하고 있습니다.

한편, 기업이 보유한 정보의 대부분이 e-mail, PDF 와 같은 문서 및 HTML 등의 비정형 텍스트 형태로 구성되어 있습니다. 이런 까닭에 질 높은 검색결과를 제공하기 위하여 비정형 문서로부터 유용한 정보를 추출, 가공하는 기술이 필요합니다. 이러한 기술을 텍스트 마이닝(Text Mining)이라고 하는데, 이때에도 역시 '한국어 정보 처리 기술'이 적용되어야 합니다.

국내 검색 전문기업들은 검색 결과의 정확도를 높이기 위하여 국어 정보화에 노력을 하고 있습니다. 특히 코난테크놀로지는 자체적으로 250만 건에 이르는 한국어 사전을 구축하고, 형태소 분석을 전담할 국문학 전공자를 채용하여 전담 팀을 꾸리는 등 우리 모국어 정보 처리를 위하여 많은 투자를 아끼지 않고 있습니다.

**<u>코난테크놀로지 회사 소개</u>**
전산학 분야 중 자연어처리 연구 전문가들의 학술 모임인 KONAN(Korean Natural Language Analysis)연구 그룹에 모태를 두고 1999년 설립된 코난테크놀로지는 텍스트 및 멀티미디어 등 디지털 컨텐츠 검색 및 자산관리에 대한 연구 개발 성과를 바탕으로 지난해 연 매출 100억 원을 기록, 올해는 전년 대비 50% 이상 성장한 150억 원의 매출을 바라보고 있는 순수 소프트웨어 기업입니다.

2008년 10월 현재 150여명의 직원이 근무 중이고 이 중 70% 이상이 연구 개발 인력입니다.
다양한 검색기술 및 멀티미디어 처리 기술을 보유하고 있고 이를 패키지화하여 솔루션으로 제공하고 있습니다. 기술지원 및 구축 노하우가 뛰어나 현재 국내 검색솔루션 전문 기업 중 시장 점유율 및 매출 1위를 기록하고 있습니다.

엠파스, 싸이월드, SK11번가, GSeStore, 잡코리아, 부동산114 등의 대규모 포털 및 e-commerce 사이트의 통합 검색 서비스를 제공하고 있으며, SK텔레콤, 우리투자증권, 삼성카드, 각 방송사, 대검찰청 등 기업 및 공공기관의 홈페이지 및 내부 업무 시스템에 검색 솔루션을 구축한 바 있습니다.

현재 멀티미디어 자산관리 솔루션을 통하여 멀티미디어 분야의 검색 및 방송 시장도 적극적으로 개척하고 있고 이 분야의 고객으로는 KBS, SBS 등 국내 굴지의 방송사와 한국토지공사, CJ 홈쇼핑 등, 유수의 기업 및 공공 부문 고객을 확보하고 있습니다.