IDC
Analyze the Future

백서

컨텐츠 분석 및 고성능 엔터프라이즈

후원: IBM

Susan Feldman
David Schubmehl
Hadley Reynolds
2012년 12월

IDC의 견해

정보는 오늘날의 조직들에 있어 매우 중요한 리소스가 되었습니다. 조직들은 고객의 의견이 무엇인지, 경쟁사의 계획은 무엇인지를 파악해야 하며, 무엇보다도 모든 정보를 마이닝(mining)하여 비즈니스에 대한 전체적인 시야를 확보할 수 있어야 합니다. 이러한 중요한 정보는 텍스트(문서, 이메일 메시지), 대중 매체의 의견, 또는 소셜 포럼 등에 포함되어 있습니다. 공개된 정보와 알려져 있지 않은 정보를 모두 발견하려면 텍스트는 물론, 음성, 소리 및 이미지 등을 포함한 컨텐츠 유형을 검색 및 분석할 수 있어야 합니다. 검색 엔진과 컨텐츠 분석은 컨텐츠를 이해, 검색 및 마이닝하기 위해 개발되어 온 기술의 집합체입니다. 이러한 기술은 정형 데이터가 제공하는 "언제", "어디서", "누가", "무엇을"에 대한 정보에 덧붙여 이벤트가 발생한 "이유"나 트렌드에 대한 정보를 추가로 제공합니다.

비즈니스 필수요소

오늘날의 글로벌 비즈니스는 매 순간 모든 조직 내외에서 유통되는 수많은 데이터를 관리해야 하는 전대미문의 과제에 직면해 있습니다. IBM은 최근 글로벌 기업을 이끄는 1,700명 이상의 CEO를 대상으로 한 설문조사(Leading Through Connections, May 2012, IBM Institute for Business Value)의 결과를 발표했으며, 이 설문조사의 결과에 따르면 CEO들은 대량의 정보 흐름을 마이닝 및 분석할 수 있는 새로운 기술을 이용해 경쟁력을 얻기를 기대하는 것으로 나타났습니다.

IBM의 연구에서, 다른 기업보다 "실적이 우수한" 기업의 CEO들은 데이터를 수용하고, 이러한 데이터로부터 통찰력을 얻고, 이러한 통찰력을 바탕으로 행동하는 역량에서 차별화되었습니다. 한 CEO는 "향후 5년 동안의 중요한 생존 기술은 다른 기업보다 먼저 통찰력을 얻는 기술이 될 것입니다."라고 말했습니다.

디지털 정보로부터 통찰력을 얻기 위해 기업들은 데이터베이스 기술, 엔터프라이즈 애플리케이션, 검색 및 비즈니스 인텔리전스(BI) 등의 일반적인 소프트웨어 툴을 사용하고 있습니다. 이러한 툴들이 IT 업무 전반에 사용되고 있긴 하나, 방대하고 빠른 데이터의 흐름에서 통찰력을 얻어 조직 내 직원과 임원들이 활용하도록 제공하는 것은 여전히 기업이 풀어야 할 과제로 남아 있습니다. 실적이 우수한 조직들은 이러한 문제를 해결하기 위해 컨텐츠 분석과 같은 혁신적인 기술을 도입하고 있습니다. 이러한 혁신적인 기술은 데이터베이스 애플리케이션의 견고한 신뢰성에 민첩성과 속도를 더하며, 정보의 85%를 차지하는 비정형 정보에 대한 통찰력을 제공합니다. 비즈니스 전문가들은 이를 통해 통찰력을 얻고, 질문에 대한 답을 찾을 수 있으며, 의사결정에 이러한 통찰력을 이용할 수 있습니다.

글로벌 본사: 5 Speen Street Framingham, MA 01701 USA 전화.508.872.8200 팩스.508.935.4015 www.idc.com

컨텐츠 분석이란 무엇인가?

컨텐츠 분석 애플리케이션은 컨텐츠를 처리하여 생각, 이름, 시간, 이벤트 및 의견과 같은 정보의 구성요소를 추출합니다. 또한 원인과 영향, 인수와 합병, 약물의 부작용과 같은 구성요소 간의 관계를 파악합니다. 컨텐츠는 사진의 캡션, 블로그, 뉴스 사이트, 고객과의 대화(오디오 및 텍스트), 소셜 네트워크에서 이루어지는 논의, 얼굴, 지도, 멀티미디어 리소스, 기존의 텍스트 문서 등 다양한 유형으로 이루어집니다. 컨텐츠 분석을 위한 기술에는 개체 추출, 관계 발견, 감성 분석, 평판 관리, 트렌드 분석, 관련성, 추천, 얼굴 인식, 음성 언어 분석, 시각화 및 각 산업 분야에 초점을 둔 맞춤식 분석이 포함됩니다. 이러한 기술들은 결합되는 방법에 따라 다음과 같은 응용 분야를 위한 기반을 구축합니다.

- 콜 센터 데이터 및 기타 고객과의 의사소통을 분석하기 위한 "고객의 소리"
- 사법 활동을 지원하기 위한 범죄 및 범죄 활동의 분석 및 발견
- 광범위한 분야의 경쟁력 있는 인텔리전스
- 새로운 제품에 대한 소비자의 반응을 측정하기 위한 제품 출시 분석
- 금융 서비스, 정부 수당 관리, 보험을 비롯한 여러 분야에서의 부정행위 발견
- 헬스케어에서부터 원예에 이르기까지 다양한 분야를 위한 전문 애플리케이션

컨텐츠 분석을 이용하면 기존 소프트웨어 애플리케이션을 강화할 수 있으며, 새로운 소프트웨어 애플리케이션의 기반을 구축함으로써 텍스트를 이해하고 마이닝하여 새로운 유형의 비즈니스 인텔리전스를 얻을 수 있습니다. 기존의 비즈니스 인텔리전스 애플리케이션에서 데이터가 사용된 방식처럼, 컨텐츠 분석을 통해 추출한 요소들은 정렬 또는 비교하거나 시간의 흐름에 따라 표시할 수 있으며, 이를 통해 트렌드를 파악하고 반복적인 비즈니스 프로세스 또는 관심 주제의 현황에 대한 보고서를 작성할 수 있습니다. BI와는 다르게, 컨텐츠 분석을 이용하면 예측 가능한 행동 및 예측 불가능한 행동과 사람들 사이의 상호작용을 모니터링함으로써 예상하지 못했던 사실을 환기시킬 수 있습니다. 컨텐츠 분석을 이용하면 새로운 트렌드, 신규 시장, 또는 고객 관계에 대한 문제점을 발견할 수 있으며, 최상위 수준의 컨텐츠 분석을 이용하면 질문에 대한 답을 직접 얻을 수 있습니다. 기존의 애플리케이션과 결합했을 때, 이러한 새로운 기술 기반은 언어 이해 기능을 추가하여 비즈니스 인텔리전스, 검색 또는 콜 센터 애플리케이션을 강화 및 확장할 수 있으며, 이를 통해 생산성을 높이고 비즈니스에 대한 더욱 명확한 시야를 제공할 수 있습니다. 이러한 이유로, 검색 및 발견과 관련된 시장 분야는 도입률과 이용률이 높아지고 있습니다.

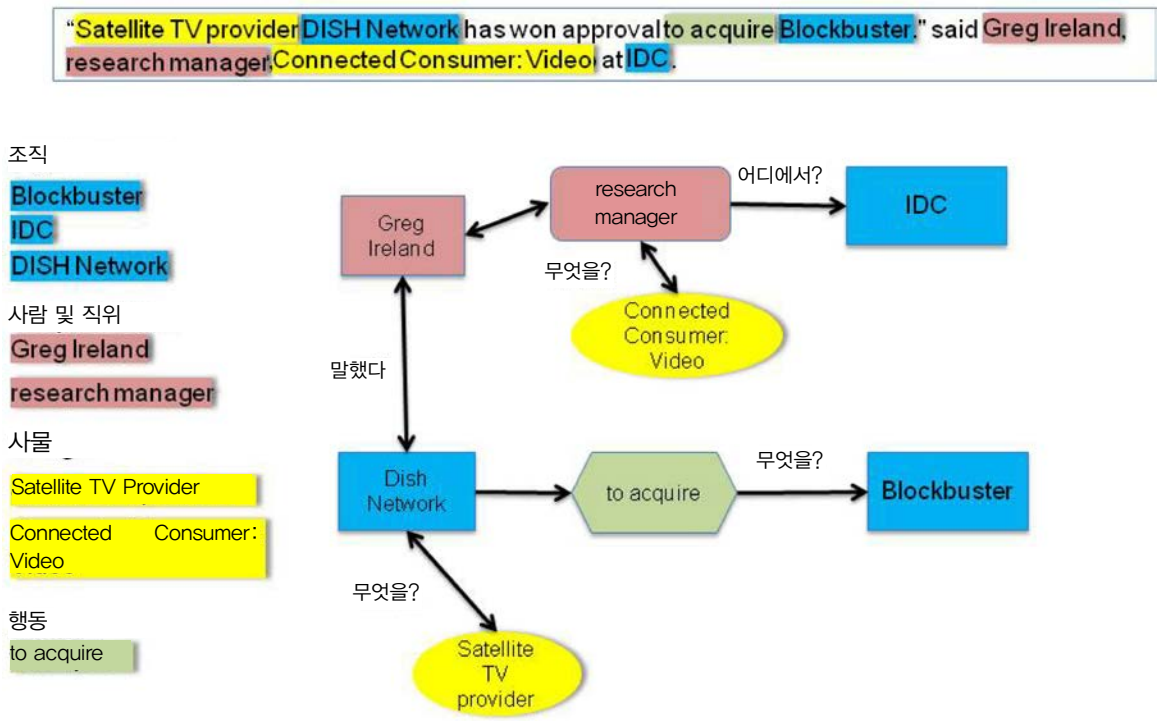
컨텐츠 분석 툴은 엔터프라이즈 검색 및 비즈니스 인텔리전스 기술을 통해 개발된 접근법을 강화하여 더욱 정확한 정보 액세스, 더욱 뛰어난 정보 집합 탐색 기능, 그리고 정형 데이터 및 비정형 데이터에 대한 통합된 시야를 제공합니다. 이는 중첩 기술(overlay technology)의 역할을 하여 검색 및 BI 애플리케이션의 성능을 향상시키기도 합니다. 이러한 기술은 통합 정보 액세스 애플리케이션의 핵심 요소이며, 이를 이용해 매우 명확하고 효율적인 의사결정 지원 환경

에 대한 기업의 필요사항이나 맞춤화된 연구 엔진에 대한 기업의 필요사항에 대해 대응할 수 있습니다.

컨텐츠 분석의 간단한 예시는 그림 1의 문장 분석에서 확인할 수 있습니다. 이 사례에서 "Blockbuster"는 블록버스터 영화가 아니라 기업의 이름이라는 것을 알 수 있으며, "Ireland"는 아일랜드라는 장소적 개념이 아니라 "IDC"라는 또 다른 조직에 소속된 연구 관리자의 이름이라는 것을 알 수 있습니다. 컨텐츠 분석은 사람, 장소, 사물 및 행동을 식별할 수 있는 수단이 되며, 여기에서 얻어진 연관성과 상관관계 정보는 엔터프라이즈 검색 시스템이나 비즈니스 인텔리전스 툴을 사용한 연구나 발견에 활용될 수 있습니다.

그림 1

컨텐츠 분석 예시



출처: IDC, 2012

컨텐츠 분석의 역할

오늘날의 빅데이터 환경은 정보 분석에 대한 복잡성을 증가시키는 동시에, 이전에는 확인할 수 없었던 고객, 환자, 시민, 트렌드와 생각을 이해할 수 있는 기회를 제공하고 있습니다. 빅데이터의 거대한 양은 분석의 확실성을 높이는 데 도움이 되며, 인구의 세분화를 통해 전자 상거래를 개인 맞춤화하거나 환자를 진단하는 데에도 사용될 수 있습니다.

조직에서 유통되는 대부분의 데이터는 텍스트, 동영상, 이미지, 소셜 미디어, 웹 페이지, 모바일 및 음성 통신과 같은 비정형 데이터입니다. 하지만, IT 부서에서 대량의 데이터를 분석하기 위해 이용할 수 있는 툴은 정보가 정확하게 행과 열로 정리되고 사전에 정의된 체계에 따라 정형화된 정형 데이터 분야를 위해 설계되어 있습니다. 데이터베이스와는 달리, 텍스트와 리치 미디어(rich media) 그리고 이러한 정보 스트림의 급격한 변화는 데이터베이스 및 데이터 웨어하우스와 같은 정형 데이터 툴로는 효과적으로 포착하고, 분석하고, 관리할 수 없습니다.

컨텐츠 분석은 비정형 정보 분야와 정형 정보 분야의 경계에 걸쳐있습니다. 컨텐츠 분석은 비정형 정보로부터 의미를 갖는 요소를 추출한 후 이러한 요소를 정형화된 형식으로 표시하며, 이러한 요소는 정형 데이터와 함께 결합 및 분석될 수 있습니다. 또한, 컨텐츠 분석을 이용하면 데이터베이스로부터 의미 있는 정보를 추출할 수 있으며, 데이터베이스와 컨텐츠 집합 모두에 걸쳐 정규화를 실행하여 소스 집합의 경계를 넘어 관계를 파악할 수 있습니다. 이러한 기술은 문서를 분류하고 태그를 지정하여 해당 문서가 무엇에 관한 문서인지 강조합니다. 컨텐츠 분석은 사람, 장소, 제품 및 사물(개체)의 이름뿐만 아니라 시간, 의견, 감성 및 지리적 위치를 추출하며, 검색 인덱스에 이러한 추가적인 정보를 메타데이터로 추가합니다. 이러한 추가적인 메타데이터는 검색 엔진이 더욱 관련성 있는 결과를 반환할 수 있도록 하여 검색 기능을 향상시킵니다. 태그(개체, 시간, 위치 및 주제)는 패싯(facet)으로 이용되며, 따라서 사용자는 검색을 하는 대신 브라우징을 통해 정보의 집합을 탐색할 수 있습니다. 컨텐츠 분석이 제공하는 정형화된 결과는 이제 기존의 BI 툴에 추가될 수 있으며, 이를 이용해 더욱 종합적인 비즈니스 분석을 실행할 수 있습니다. 트렌드 추적 소프트웨어에서는 이러한 결과를 이용해 주식, 질병 패턴 또는 판매에 대한 트렌드를 발견할 수 있습니다. 컨텐츠 분석을 이용하면 질문에 대한 답을 자동으로 제공하는 기능을 온라인 셀프 헬프 사이트에 추가할 수 있으며, 비정형 정보에서 발견된 긍정적이거나 부정적인 정보들을 임원진에게 제공할 수 있습니다. 고객의 불만사항을 마이닝하면 심각한 문제를 식별하고 제품과 관련된 문제의 발생을 방지할 수 있습니다. 컨텐츠 분석은 eDiscovery 및 제품 개발에 이용될 수 있습니다. 즉, 컨텐츠 분석을 이용하는 것은 분석 분야에 있어서 퍼즐의 빠진 한 조각을 찾는 것과 같습니다.

컨텐츠 분석 작업은 정보의 정제 작업과 연계되어 왔으며, 어노테이터(annotator), 추출 툴(extractor) 또는 인식 툴(recognizer)로 불리는 인텔리전스 처리기(intelligent processor)를 통해 새로운 수준의 의미와 관점을 원본 텍스트 및 인덱스에 추가할 수 있습니다. 이러한 새로운 데이터 요소는 거의 무한대의 의사결정 분야에 대한 일련의 새로운 렌즈를 제공하며, 비즈니스 구현 측면에서는 이러한 요소를 자신의 특정한 관심 분야에 맞도록 조정할 수 있습니다. 임원진은 비즈니스 의사결정을 위해 점점 더 많은 정형 소스 및 비정형 소스의 데이터를 검토하도록 요구받고 있으며, 컨텐츠 분석은 이러한 정형 정보와 비정형 정보를 서로 연결합니다. 컨텐츠 분석은 먼저 분석 작업에 비정형 정보를 제공하며, 관리자는 이를 이용해 모든 관련 정보에 대한 시야를 통합하여 분석, 발견 및 의사결정에 이용할 수 있습니다.

대부분의 조직은 텍스트를 다루기 위한 첫 번째 단계를 수행하기 위해 검색 엔진을 배치하고 있습니다. 그러나 검색 작업을 통해 질문에 대한 답을 얻을 수는 있지만 쿼리를 실행하지 않으면 정보를 탐색할 수 없으므로, 검색만으로는 더욱 발전된 분석을 실행할 수 없습니다. 정보의 탐색을 통해 예상치 못했던 결과를 얻을 수도 있으며, 이러한 정보는 정보 마이닝 작업의 가장 값진 결과라고 할 수 있습니다. 엔터프라이즈 검색 기능을 설치하는 것만으로도 비정형 데이터의 집합과 흐름을 더 잘 이해할 수 있지만, 데이터 및 텍스트 집합에 대한 투자 수익률을 최대화하고 비정형 정보 및 정형 정보 리소스로부터 가장 큰 가치를 얻으려면 컨텐츠 분석 기능을 추가해야 할 것입니다. 지능적인 컨텐츠 분석을 이용하면 정보 시스템에서 다음과 같은 작업을 실행할 수 있습니다.

- 텍스트와 기타 비정형 및 정형 정보의 정보 구성요소를 분류 및 분석하고 정보 구성요소에 태그를 지정
- 개체, 이벤트 및 관계를 통해 정보를 발견 및 탐색
- 더욱 정확하고 맥락에 맞는 검색 결과를 제공
- 밑바탕이 되는 컨텐츠 및 데이터 소스에 대한 트렌드를 BI와 유사한 방식으로 시각화하여 제공

컨텐츠 분석이라는 분야가 강력하고 체계적인 방식이라는 것은 분명한 사실이지만, 대부분의 기업에게 이는 아직 새로운 분야입니다. 컨텐츠 분석을 기반으로 하는 다양한 애플리케이션이 더욱 널리 도입되면서 컨텐츠 분석에 대한 투자는 급속히 증가하고 있습니다. 개념, 개체 및 기타 메타데이터를 추가하여 검색 기능을 향상시키는 컨텐츠 분석은 "고객의 소리" 애플리케이션, eDiscovery, 제약회사의 연구 및 정부의 정보 기관 등에서 널리 이용되고 있습니다. 금융 업계는 부정행위를 발견하고 고객에게 투자를 추천하기 위해 컨텐츠 분석을 도입하고 있습니다.

컨텐츠 분석 소프트웨어는 엔터프라이즈 검색 등의 기술보다 큰 가치를 즉각적으로 제공합니다 (예: 검색 결과가 첨부되는 탐색 인터페이스를 생성하는 경우). 큰 가치를 제공하는 복잡한 활용을 위한 엔터프라이즈급 컨텐츠 분석 업무에는 복잡한 기술들로 이루어진 툴박스가 필요하며, 조직들은 이러한 툴박스를 비즈니스 환경, 그리고 제품 및 서비스와 관련된 정보 영역에 맞게 조정하고자 합니다. 비즈니스를 더욱 정확하게 이해하려는 경우, 프로세스를 시작하고 행동을 개시하기 위해 분류 체계, 범주 및 규칙과 같이 비즈니스 목표 및 의사결정 환경과 관련된 특정 용어를 추가해야 할 필요가 있습니다. 쉽게 액세스할 수 있는 효율적인 컨텐츠 분석 모델링 환경은 실적이 우수한 조직들에서 볼 수 있는 기본요소입니다. 이러한 툴들을 이용하면 새로운 수준의 혁신적인 분석 애플리케이션을 개발할 수 있으며, 컨텐츠 분석을 통해 여러 가지 업무를 추진할 수 있습니다.

컨텐츠 분석 실제 사례:
컨텐츠 분석의 잠재력과 위험

조직은 변화를 수용해야 합니다. 이는 비즈니스에 영향을 미치는 새로운 트렌드의 발견, 이러한 트렌드가 미치는 영향의 특성에 대한 이해, 변화에 대한 전략의 수립 및 변화 관리 계획의 수립이 신속하게 이루어져야 한다는 것을 의미합니다. 컨텐츠 분석은 기존의 툴에 비해 향상된 기능들을 제공하며, 특히 발견 및 이해와 관련된 분야에서는 더욱 그렇습니다. 하지만 이러한 새로운 시스템을 구현하는 데는 고유한 과제가 존재합니다. 이 절에서는 컨텐츠 분석 애플리케이션이 제공하는 결과의 품질에 영향을 미칠 수 있는 여러 가지 역학관계를 검토하고, 현재 컨텐츠 분석을 성공적으로 이용하고 있는 조직에서 발견한 주요 사례를 살펴봅니다.

텍스트는 데이터가 아닙니다

텍스트가 포함하고 있는 정보 유형의 관점이나 텍스트의 형식의 관점에서 보면 텍스트는 데이터가 아닙니다. 데이터에는 어떠한 이벤트에 대한 "무엇을, 어디서, 언제"의 항목이 포함되지만, 텍스트에는 "왜"와 "어떻게"의 항목이 포함되어 있는 경우가 많습니다. 텍스트는 모호하고 예측할 수 없으며, 동일한 생각을 여러 가지 방법으로 서술할 수 있습니다. 컨텐츠 분석은 숫자가 아닌 언어를 분석하도록 설계되었으며, 결정론적이기보다는 확률론적인 기술입니다. 컨텐츠 분석은 의미의 매칭을 통해 동의어를 수용하고 텍스트의 표현을 재구성합니다. 이러한 통계적인 유연성 덕분에 컨텐츠 분석에서는 단순히 데이터베이스 항목이나 검색어와 정확히 일치하는 표현을 발견하는 것이 아니라, 유사한 의미를 내포하고 있는 표현을 발견할 수 있습니다.

컨텐츠 분석 툴에는 기존의 데이터베이스나 데이터 웨어하우스 애플리케이션에서 이용되는 것과는 다른 종류의 처리가 필요합니다. 정형 데이터 분야에서, 데이터는 사전에 결정된 질문에 대한 답을 제공하도록 사전에 설계된 개요에 따라 정렬됩니다. 텍스트에 대한 퍼지(fuzzy) 분야에서, 개념, 이름, 이벤트, 관계, 감성, 위치, 시간과 같은 귀중한 정보는 유연성 있는 인덱스로 추출되어 사용자가 검색을 실행하거나 질문을 할 때 결합되고 표현될 수 있습니다. 일반적으로 컨텐츠 분석 애플리케이션은 텍스트에 포함되어 있는 개체, 관계 또는 상관관계를 발견할 수 있는 기회를 제공하며, 이러한 요소는 이전에 정형화된 데이터에서 발견할 수 없었던 요소일 수 있습니다.

정형 데이터 분야에서, 비즈니스 인텔리전스 솔루션은 데이터 내의 연상 패턴 및 관계를 발견하는 데 점점 더 많이 이용되고 있습니다. 텍스트와 같은 비정형 데이터 분야에서 컨텐츠 분석은 발견과 분석을 위한한 핵심 요소이며, 사전에 정의된 개요를 이용하지 않고도 퍼지 매칭 및 기계 학습과 같은 확률론적 툴을 이용해 패턴을 식별하고 관계 고리를 설정할 수 있도록 합니다.

데이터가 증가할수록 선택의 폭이 넓어집니다

데이터는 더 많을수록 좋습니다. 데이터의 변화가 트렌드를 나타내는 것인지 무시해도 되는 것인지를 판별하려면 더 많은 데이터를 수집하여 추측을 수행해야 합니다.

5명이 애완용 제품을 구매할 때와 50,000명이 애완용 제품을 구매할 때의 차이를 확인하는 것은 생산 및 유통과 관련된 의사결정을 하는 데 도움이 될 수 있습니다. 많은 양의 데이터는 정보 분야의 얼리 어답터 기업에게 다른 기업보다 먼저 트렌드나 위협을 발견할 수 있는 이점을 제공합니다.

비정형 데이터는 특별한 과제를 제시합니다. 기업 내의 모든 데이터와 웹상의 모든 데이터의 85% 이상은 비정형 데이터이며, 그 동안 이러한 데이터는 정형 데이터 지향적인 비즈니스 분석 프로세스에서 불분명한 존재였기 때문입니다. 조직들은 처리해야 하는 데이터의 양을 줄이기 위해 전략을 수립하려 해 왔으며, 데이터의 지속적인 팽창을 이점이 아닌 위협으로 간주했습니다.

컨텐츠 분석은 대량의 데이터를 기업의 이점으로 승화시키는 솔루션입니다. 이 솔루션의 작동 원리는 텍스트 문서 형식의 데이터를 데이터베이스 친화적인 형식으로 표현하는 것이며, 기존의 비즈니스 분석 솔루션은 이러한 데이터를 판독 및 활용하여 의사결정을 위한 종합적인 통찰력을 제공합니다.

특히, 고객 정보와 관련된 컨텐츠 분석은 비즈니스에 대한 통찰력을 제공하는 데 큰 공헌을 하고 있습니다. 오늘날의 소셜 미디어 환경은 빅데이터와 관련된 과제를 던져주었으며, 지속적으로 증가하는 웹상의 소셜 데이터 집합에서 사용자는 끊임 없이 기업과 기업의 제품에 대한 의견을 제공하고 있습니다.

컨텐츠 분석에 대한 준비

컨텐츠 분석 툴을 이용해 텍스트 데이터를 정리하십시오

컨텐츠 분석을 시작하기에 어느 정도의 정리가 충분한지 결정하십시오

완전히 깨끗한 데이터는 없습니다. 즉, 데이터는 관리하고, 정리하고, 정규화해야 합니다. 비정형 컨텐츠 분석 분야에서, 오염된 데이터는 분석 및 검색 프로세스 도중에 불완전한 답을 제공할 수 있습니다. 그 이유는 다음과 같습니다.

- 철자법이 잘못되었거나 및 대안적인 철자법이 이용되는 경우
- 모호한 참조를 이용하는 경우
- 전문 용어가 변경되는 경우
- 실제로 오류가 발생하는 경우
- 데이터를 더 이상 사용할 수 없는 경우
- 데이터가 부정확한 경우

컨텐츠 분석 구현 팀이 첫 번째로 실행해야 할 작업은 컨텐츠 분석 소프트웨어가 제공하는 데이터 품질 관련 문제 해결 툴을 학습하는 것입니다. 데이터의 정리를 완전하게 수작업으로 진행해서는 안 됩니다. 고급 컨텐츠 분석 솔루션은 데이터 품질 관련 문제를 처리할 수 있는 툴을 제공합니다. 이러한 툴은 잘못된 철자 및 대안적인 철자(Alternative spelling), 모호한 참조 등을 발견합니다. 또한, 오염되거나 불완전한 데이터를 효율적인 방식으로 발견 및 처리할 수 있도록 다수의 통계적인 방법 및 사람에 의한 일반적인 테스트 및 검토 절차를 제공합니다.

텍스트 데이터는 100% 정확할 수 없다는 사실에 유의해야 합니다. 데이터 품질에 대해 잘 알려지지 않은 비밀은, 수작업으로 입력하거나 정리한 데이터일지라도 그 정확도는 겨우 70%~80%에 그친다는 것입니다. 자동화된 분류 및 태그 지정 기법의 정확도는 일반적으로 최소 70%~80% 수준이며, 시간이 흐름에 따른 맞춤화를 통해 정확도가 90%~95%까지 향상될 수 있습니다. 컨텐츠 분석 기법이 100% 정확한 것은 아니지만, 대부분의 조직들이 정보에 대한 통찰력을 전혀 확보하지 못 하고 있다는 것을 감안하면 80%의 정확도만으로도 크게 발전된 수준이라고 볼 수 있습니다. 조직은 유용한 결과를 제공하는 데 필요한 최소한의 정확도 수준을 결정해야 합니다

고객 관계 및 마케팅 담당 관리자는 고객 정서 모니터링 기능을 제공하는 컨텐츠 분석 소프트웨어를 이용하여 콜 센터에 접수된 통화뿐만 아니라 지원 사이트에 접수된 이메일, 고객과의 채팅, 고객 설문조사 양식에 대한 반응, 그리고 Facebook 및 Twitter와 같은 서비스에 게시된 소셜 미디어 게시물을 마이닝하여 고객 정서에 대한 통합된 시야를 확보할 수 있게 되었습니다. 이는 분석 기술의 발전을 활용하는 기업에 빅데이터가 제공하는 통찰력을 보여줍니다.

분석 영역이 컨텐츠 분석 품질에 미치는 영향

식료품점의 직원에게 흉부 엑스선 사진에 대한 분석을 맡기는 일이 일어나서는 안 될 것입니다. 즉, 컨텐츠 분석 시스템에 기업, 산업 분야, 제품 및 인물에 대한 지식을 입력하는 것은 가장 중요한 일입니다. 애플리케이션의 환경, 일반적인 분야 또는 영역, 비즈니스의 특정한 유형에 속하는 산업 분야나 전문 분야의 용어, 개념 및 지식을 갖추고 있는 경우, 컨텐츠 분석 애플리케이션의 정확성은 향상됩니다. 조직의 상황에 대한 구체적인 정보는 해당 컨텐츠 내에서 가장 관련성이 높은 문제의 종류, 문제, 용어 및 개념적 관계를 규명하는 데 중요한 역할을 합니다.

해당 영역에 대한 맥락은 컨텐츠 분석 시스템의 정확도 및 영향력에 있어 매우 중요한 역할을 합니다. 예를 들어, 헬스케어 분야의 환경에서 이용되는 애플리케이션에는 항공기 유지보수 분야에서 이용되는 컨텐츠 분석 애플리케이션과는 전혀 다른 메타데이터, 마크업, 분석 및 표현이 필요할 것입니다. 헬스케어 영역 내에서도, 최전선의 간병인에게 서비스를 제공하는 컨텐츠 분석 애플리케이션은 병원 관리 또는 보험 업계의 보험료 납부자 측 담당자를 위해 설계된 컨텐츠 분석 애플리케이션과는 크게 다를 것입니다.

산업 분야, 조직의 구조, 전문 직종 및 특정한 작업의 맥락에 따라 특정 용어와 개체 정의, 구조 및 관계가 필요할 수 있습니다. 컨텐츠 분석 애플리케이션의 개발자는 사용자가 자신의 역할과 애플리케이션의 작동 방법을 알 수 있도록 하나의 시스템 설계에 이러한 모든 요소를 통합할 수 있습니다. 이에 반해, 일반적인 시스템은 작업자가 다른 방법을 이용하거나 결과를 재해석하도록 요구합니다.

컨텐츠 분석 소프트웨어를 이용하면 애플리케이션 개발자는 컨텐츠로부터 의미를 갖는 핵심 요소를 추출하여 해당 영역의 전문 지식을 사전, 분류, 또는 다른 지식 베이스의 형태로 입력할 수 있습니다. 개발자는 다음과 같은 접근법을 사용할 수 있습니다.

- 데이터 내의 개념적, 개체 지향적, 또는 언어적 패턴을 노출시키는 자동 클러스터링 루틴
- 해당 영역의 지식 구조를 반영하여 데이터 내의 항목을 분류하는 자동 분류 및/또는 규칙 중심의 분류
- 해당 영역의 특정적인 의미를 갖는 용어와 구절을 구별하기 위한 용어 목록, 어휘 목록 및 동의어 사전
- 제품, 지역, 사무소, 딜러, 파트너, 경쟁업체 또는 해당 영역 내의 특별한 개체 및 관계에 대한 목록

- 합병 및 인수, 최초의 주식 공모, 합작 투자, 질병, 해독제, 부작용의 원인, 발명가 및 특허, 범죄자 및 범죄와 같은 관계의 유형에 대한 상세 설명

업계 전문가가 전문적인 경험을 통해 가치 있는 지식을 습득하고 더 지혜로워지는 것처럼, 이 모든 기능은 정보 시스템이 더 똑똑한 시스템으로 거듭나 데이터의 의미를 더 잘 이해할 수 있도록 합니다.

컨텐츠 분석 구성요소

컨텐츠 분석 모듈

컨텐츠 분석은 여러 요소로부터 의미 및 구조를 추출하는 일련의 모듈을 사용하며, 이를 통해 인간이 이해할 수 있는 텍스트를 생성합니다. 이러한 모듈은 각각 별도로 이용되어 정보의 검색 가능성과 발견 가능성을 향상시킵니다. 일반적인 모듈에는 다음이 포함됩니다.

- 언어 분석기는 어간을 발견하여 해당 텍스트에 품사에 대한 태그를 지정합니다.
- 개체 어노테이터(Entity annotator)는 사람, 장소 또는 사물의 이름을 인식합니다.
- 개념 어노테이터(Concept annotator) 또는 분류기(Categorizer)는 각 문장, 문단 및 문서의 주요 주제를 판별합니다.
- 관계 어노테이터(Relationship annotator)는 설명되는 행동, 개체 사이에 존재하는 관계, 그리고 작용하거나 적용되는 개념의 유형을 판별합니다.
- 날짜 어노테이터(Data annotator)는 어떠한 행동이 일어나는 때가 언제인지 판별합니다.
- 정서 어노테이터(Sentiment annotator)는 표현된 의견이 특정한 개체에 대해 호의적인지 또는 비호의적인지 판별합니다.
- 지리적 어노테이터(Geographic annotator)는 지도의 좌표를 텍스트 내의 지리와 관련이 있는 모든 개체나 개념과 연관시킵니다.

제품 구입 후 즉시 이용 가능한 어노테이터의 목록은 지속적으로 확대되고 있습니다. 그러나 애플리케이션을 특정 용도 및 컨텐츠 변화에 맞춰 조정하려면 구현 작업에서 하나 이상의 맞춤식 어노테이터를 개발해야 할 것입니다.

컨텐츠 분석 파이프라인

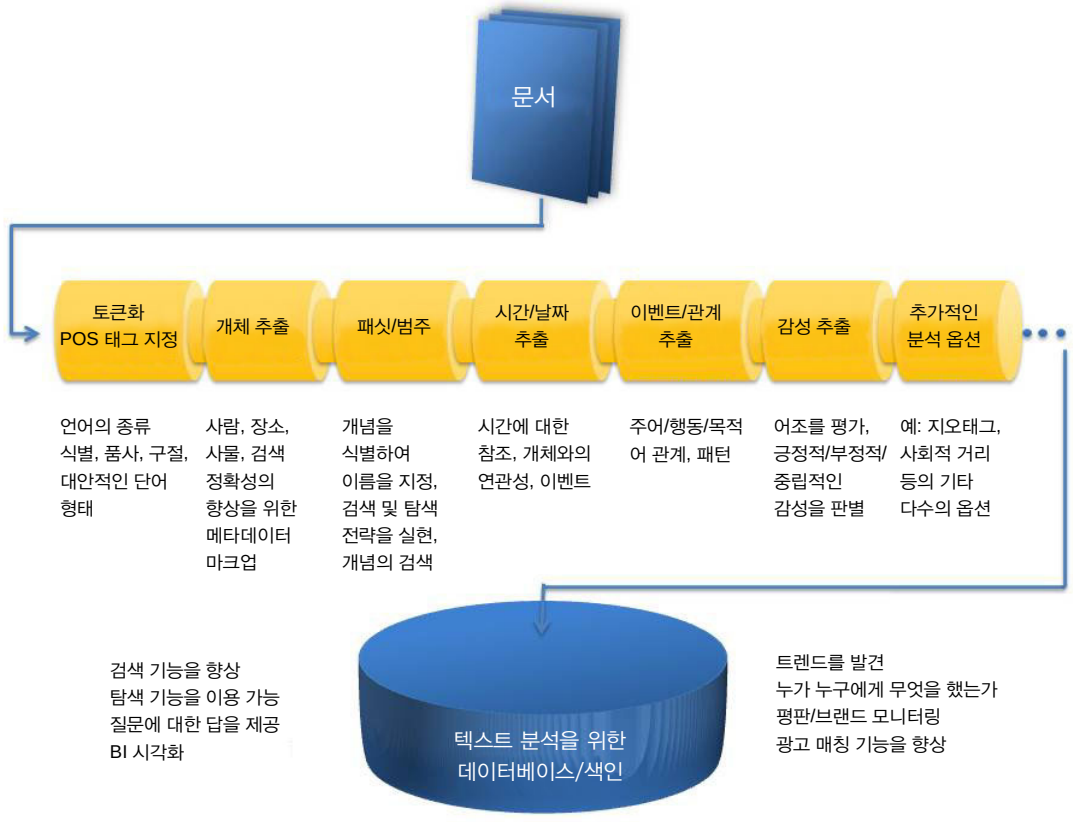
컨텐츠 분석 처리에서, 소프트웨어 제품 공급업체 및 구현 작업자에게 가장 현실적인 접근법은 파이프라인 아키텍처인 것으로 증명되었으며, 이는 그림 2에 표시되어 있습니다. 특정한 애플리케이션 내에는 많은 종류의 구성요소가 뒤섞여 있으며, 대부분의 구성요소는 작동을 위해 동일한 컨텐츠 분석 "기본요소(primitives)"를 요구합니다. 파이프라인 접근법에서는 논리적이고 효율적인 진행 과정의 태그 지정 및 추출 프로세스를 통해 컨텐츠를 단계별로 분석합니다. 첫 번째 단계에서는 이후의 단계에서 이용하게 되는 중요한 언어적 데이터(예: 언어의 종류, 품사)를 규명하며,

이후의 단계에서는 이전에 추출된 데이터(예: 감성 분석)를 요구하는, 높은 수준으로 특성화된 루틴으로 분기할 수 있습니다.

각각의 컨텐츠 분석 구성요소는 구현 전체에 고유한 가치를 더하며, 최종 제품은 검색 기능, 컨텐츠 분석 또는 비즈니스 분석 애플리케이션의 형태로 제공될 수 있습니다. 본 문서의 이후의 절에서는 컨텐츠 분석 파이프라인의 주요 요소를 확인하고 각 요소의 장점을 알아봅니다. 각 절의 순서는 현재 이용되고 있는 애플리케이션 내의 일반적인 처리 순서를 반영하였습니다.

그림 2

컨텐츠 분석 파이프라인



출처: IDC, 2012

언어 인식, 토큰화 및 형태적 분석

언어의 종류를 식별하는 것은 컨텐츠 분석 파이프라인의 첫 번째 단계입니다. 문서가 어떠한 언어로 작성되었는지 알지 못하면 그 의미를 이해하기 어렵습니다. 컨텐츠 분석 또는 검색 애플리케이션은 언어 종류 식별, 토큰화(글자들을 단어 단위로 구분) 및 어간의 식별과 같은 작업을 실행한 후 더욱 고급화된 분석 단계로 진행합니다. 애플리케이션은 다양한 언어를 처리할 수 있어야 합니다. 전 세계 각양각색의 언어로 된 컨텐츠를 처리하기 위해서는 컨텐츠 분석 소프트웨어를 도입하기 전에 이 기능을 반드시 구현해야 합니다.

구문 분석

언어적인 분석은 모든 컨텐츠 분석 시스템의 핵심적인 구성요소입니다. 이 구성요소는 구문 또는 문법의 수준에서 언어를 분석하며, 문장 내 각 단어의 역할을 발견합니다. 명사 및 동사와 같은 기초적인 형태만을 식별하는 컨텐츠 분석 애플리케이션도 있고, 단어를 어근 및 여러 가지 어미로 분석하고 텍스트 내의 모든 용어의 품사를 인식하고 구절을 식별하여 이를 "문장 도식화" 과정의 올바른 위치와 연관시키는 애플리케이션도 있습니다. 이러한 처리의 결과는 집합 내 모든 문서에 대한 메타데이터 마크업으로 나타납니다. 구절은 패싯 탐색에 이용될 수 있습니다. 컨텐츠 분석 애플리케이션이 한 문장을 완전히 분석한 후에는, 개체 사이의 관계를 이용해 원인, 효과 또는 용어의 정의를 결정할 수 있습니다. 이러한 심화 구문 분석은 이벤트 추출을 위한 필수 요소입니다.

시맨틱 분석

시맨틱 분석은 단어나 구절의 의미를 규명합니다. 이 단계에서는 소프트웨어의 버그(bug)가 실제 벌레와는 다르다는 것을 인식하는 것처럼, 모호한 용어와 관련된 문제가 해결됩니다. 함께 쓰이는 단어에 대한 사전과 패턴을 이용하면 의미를 규명하는 데 도움이 됩니다. 시맨틱 분석은 분류 작업에 도움을 주며, 분류 작업 또한 시맨틱 분석에 도움을 줍니다. 시맨틱 분석은 검색 결과를 더욱 정확하게 만들며, 분석 시각화에 대해 더욱 정확한 입력을 제공하여 컨텐츠 분석을 향상시킵니다.

실제 사례 소개

컨텐츠 분석 처리 표준

표준을 기반으로 하는 컨텐츠 분석 소프트웨어를 활용하십시오

거의 모든 컨텐츠 분석 애플리케이션은 독자적인 API를 이용해 기능을 통합합니다. 따라서, 많은 조직에서는 다양한 벤더의 애플리케이션을 통해 통합 분석을 구성하는 데 큰 어려움을 겪어 왔습니다. 이러한 비호환성을 처리하기 위해 표준을 확립하기 위한 작업이 진행되어 왔습니다. 2005년에 IBM은 미국 정부의 지원을 통해 기업 최초로 UIMA(Unstructured Information Management Architecture)를 개발하였습니다. 2006년에는 UIMA를 위한 코드를 오픈 소스 프로세스로 공개했으며, 이는 현재 Apache Software Foundation에서 관리되고 있습니다. 후속 오픈 소스 작업들은 Apache의 OpenNLP, University of Sheffield의 GATE 또는 Apache Lucene/Apache Solr 배포판과 같은 자체적인 "표준"을 제공하고 있습니다. 많은 조직들이 이러한 시스템과 UIMA를 이용하고 있습니다. 컨텐츠 분석 팀은 애플리케이션 개발 프로젝트에서 공개 표준을 기반으로 하는 툴을 이용하여 독자적인 형식의 통합과 관련된 문제의 발생을 피해야 합니다.

사전/어휘 목록

이전에 설명한 것과 같이, 영역적 맥락은 정확한 분석에 매우 중요한 요소이며, 컨텐츠 분석을 기반으로 하는 애플리케이션의 유용성에도 매우 중요합니다. 산업 분야, 조직의 구조, 전문 직종 및 특정 작업의 맥락에 따라 특정 영역의 용어와 개체 정의, 구조 및 관계가 필요할 수 있습니다. 컨텐츠 애플리케이션 개발자는 컨텐츠 분석 시스템의 사전 구성요소를 이용해 자신만의 용어 목록을 규정하고 해당 애플리케이션의 맥락에서 이러한 용어의 중요성을 정의할 수 있습니다. 이후 이러한 용어, 동의어 사전 또는 통제된 어휘를 제공하여 초기의 인덱싱 루틴에서부터 쿼리 응답 및 관련성 순위 설정에 이르는 다양한 시스템 기능에 대한 지침으로 이용할 수 있습니다. 검색 인터페이스는 동의어 사전을 통해 최상위 용어를 규명하고 용어의 폭을 좁혀줍니다.

명칭이 있는 개체의 인식

컨텐츠 분석 시스템에는 명칭이 있는 개체를 인식하는 구성요소가 있으며, 이는 사람, 장소 및 사물에 대한 논리적 정의와 물리적 정의를 인식하여 활동 및 이벤트의 분석에 대한 기초를 제공합니다.

그림 3A부터 시작되는 연속적인 스크린 샷에 나타난 것과 같이, IDC 연구 문서 저장소에서 "ireland"에 대한 표준 검색을 하면 약 4,000건의 결과가 표시됩니다. 그러나 사용자가 IDC 동영상 분석가 Greg Ireland에 의한 비디오게임 보고서를 찾고 있는 경우, 사용자는 IBM Content Analytics with Enterprise Search 소프트웨어의 좌측 패널의 "Individual(개인)" 어노테이터에서 명칭이 있는 개체에 대한 인식 기능을 이용할 수 있습니다. "Individual" 에서 Greg Ireland의 이름을 클릭하면 결과 목록을 좁힐 수 있으며, 196개의 문서가 Greg과 연관되어 있다는 것이 표시됩니다(그림 3B). "Document Cluster(문서 클러스터)" 어노테이터 섹션에서 "videogaming(비디오게임)"을 한 번 더 클릭하면 Greg이 비디오게임과 관련하여 171건의 문서를 작성했다는 것이 표시되며, 가장 관련성이 높은 문서인 2011년 11월 IDC Web Conference의 PowerPoint 컨퍼런스 발표 자료에 대한 썸네일이 표시됩니다(그림 3C).

실제 사례 소개

영역별 리소스의 활용

영역에 리소스를 식별하고 이용하기

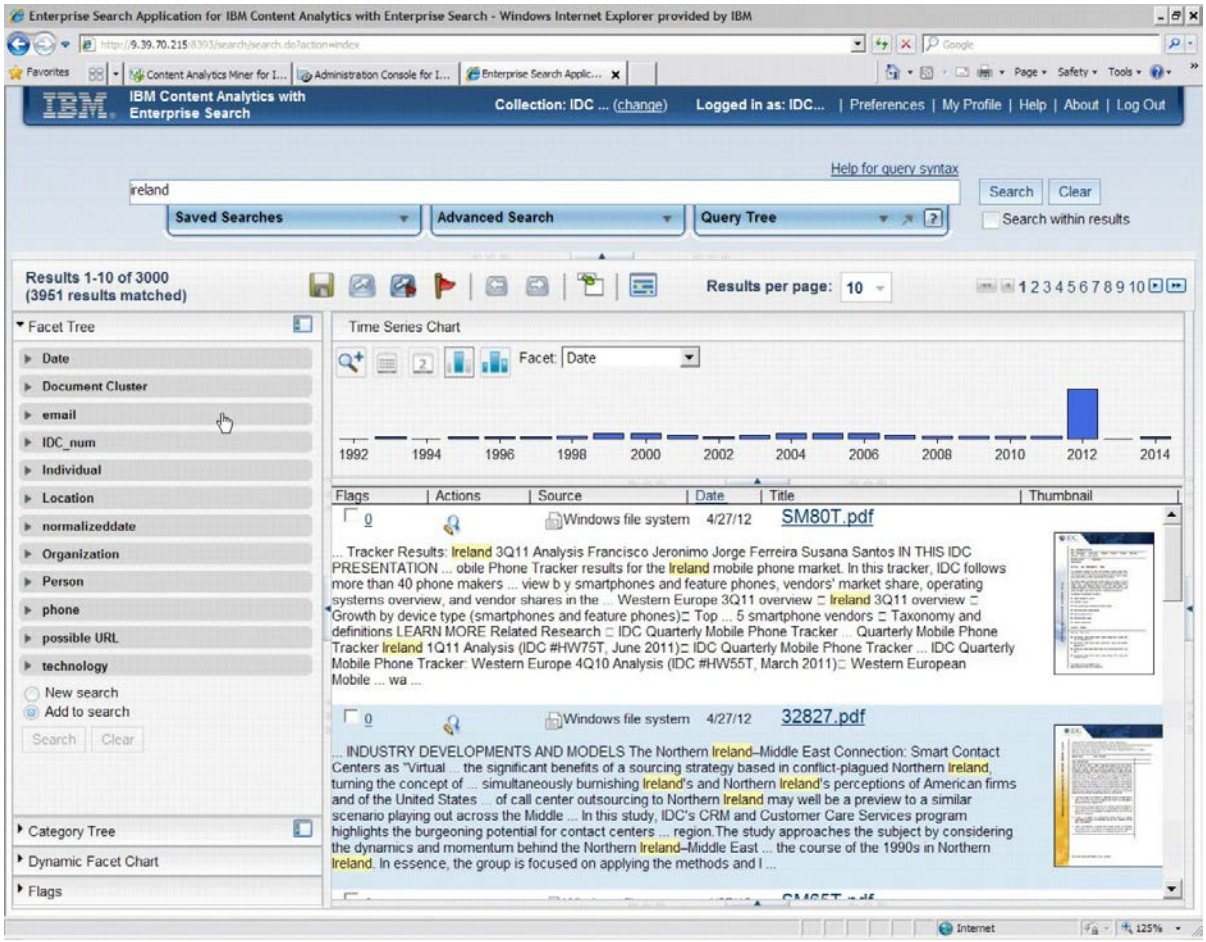
컨텐츠 분석 애플리케이션 개발자가 이용할 수 있는 영역별 리소스는 놀라울 정도로 많으며, 이러한 리소스의 대부분은 무료로 제공됩니다. 예를 들어, 미국 내 다수의 연방 정부 기관은 해당 기관의 활동과 관련된 분류를 공개관리하고 있습니다. 또한, 업계의 협회에서도 매우 상세한 분류 또는 참고문헌을 공개하고 있습니다. 예를 들면, 미국의 National Library of Medicine에서 제공하는 의학 분야의 분류인 MeSH는 생의학 관련 문헌(생명 과학 및 보건 관련 저널의 기사)을 위한 인용 자료의 참고문헌 데이터베이스인 MEDLINE에 대한 메타데이터 지침으로 이용되고 있습니다. WordNet은 여러 언어로 동의어 사전을 제공하고 있으며, 단어의 의미를 명확히 알려줍니다. 미국 증권거래위원회(SEC)에서 상장 기업의 재무 보고를 위해 도입한 XBRL 형식과 같은 정형 문서 표준은 금융 분야의 컨텐츠 분석 개발자에게 재무 데이터를 분석할 수 있는 툴을 제공할 수 있습니다.

상용 소프트웨어 업체로 구성된 한 소규모 그룹에서는 다양한 산업 분야 또는 프로세스에 대한 분류 모음을 제공하고 있습니다. 한 기업의 조직구조, 비즈니스 라인 체계, 또는 인사 부서에서 제공하는 직원 역할의 정의는 컨텐츠 분석 애플리케이션에 이용할 수 있는 잠재적인 리소스를 제공합니다.

컨텐츠 분석 애플리케이션 개발자는 이러한 리소스를 인지하고 있어야 하며 리소스로부터 최상의 가치를 얻기 위한 창의적인 방법을 생각해야 합니다.

그림 3A

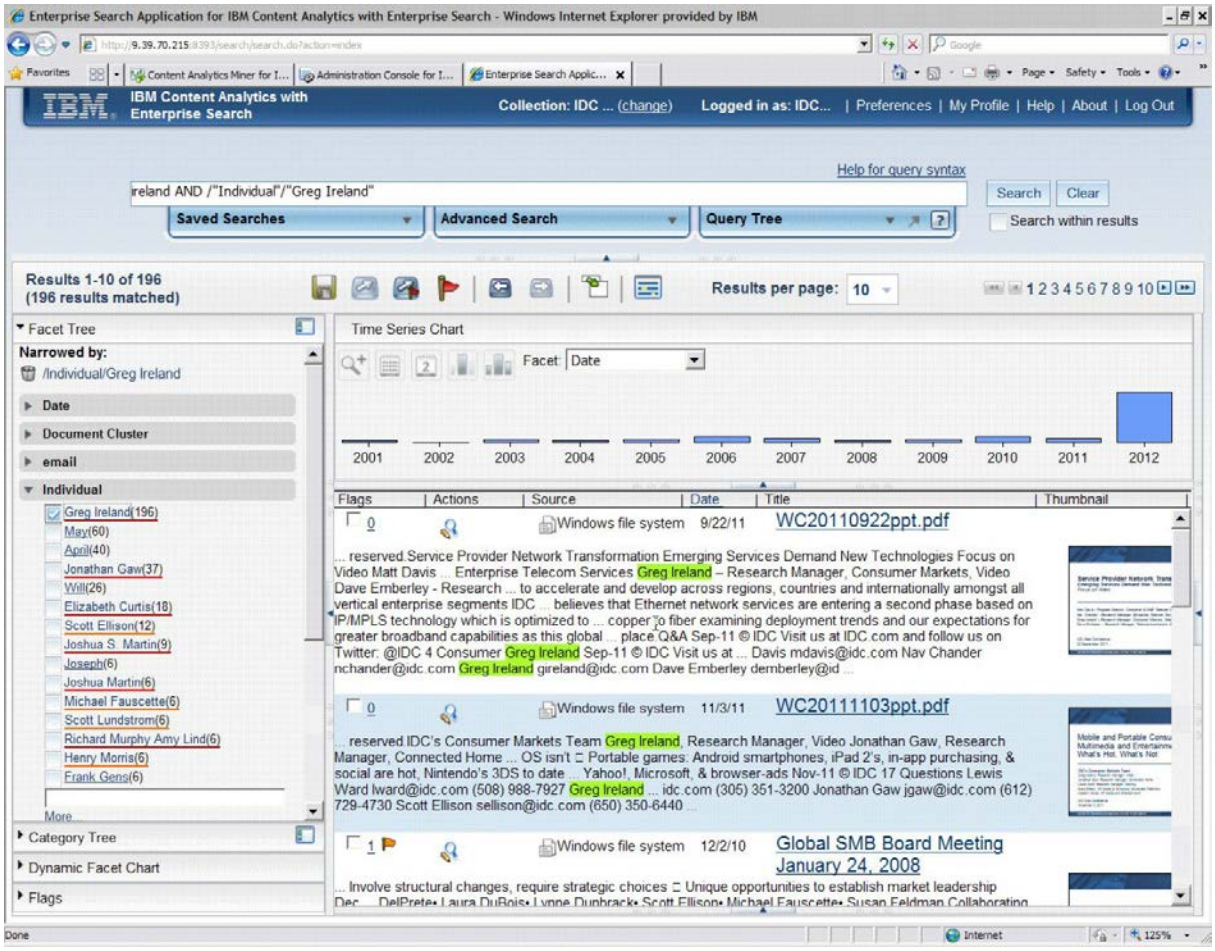
"Ireland"에 대한 검색 예시



출처: IBM, 2012

그림 3B

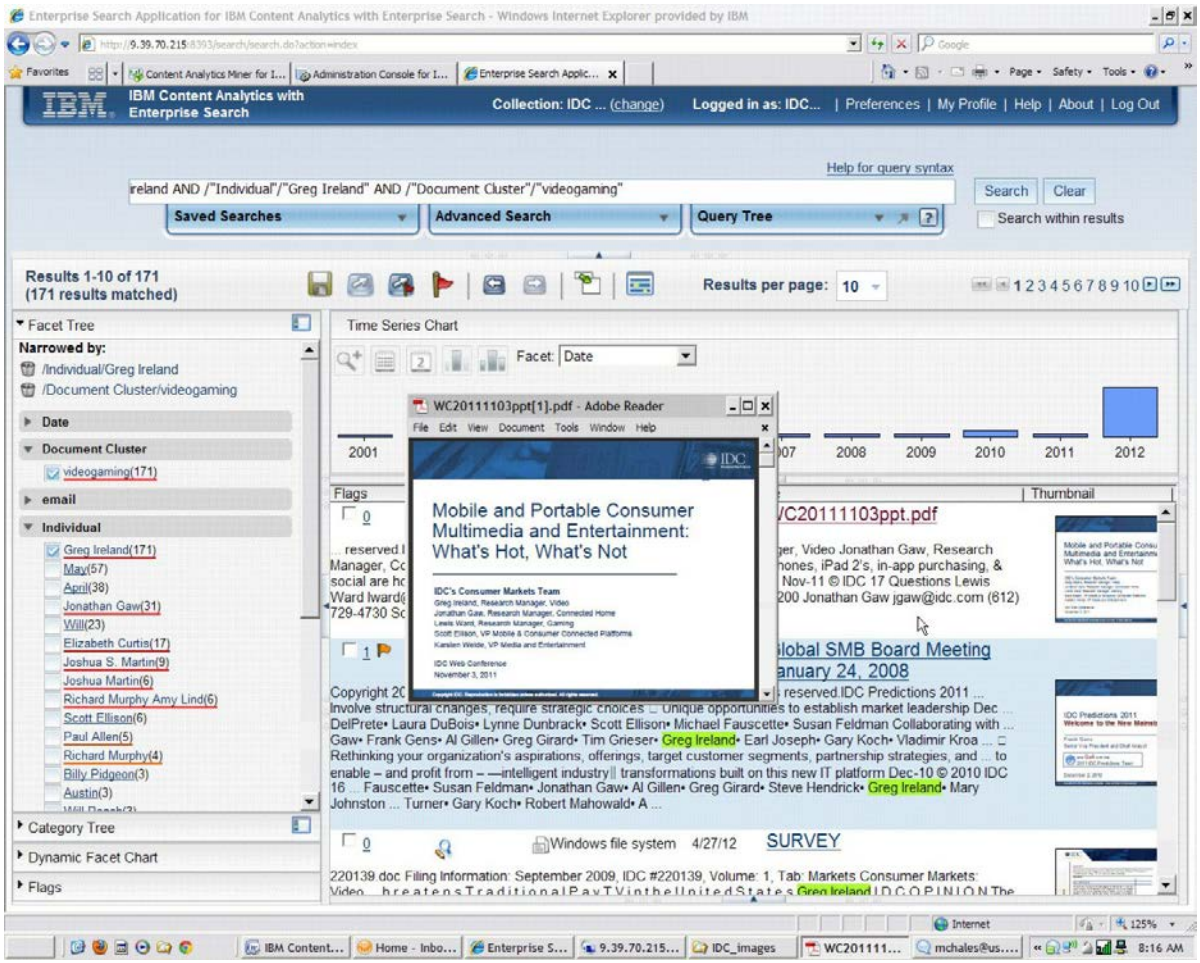
"Ireland"라는 개인에 대한 검색 예시



출처: IBM, 2012

그림 3C

문서 미리보기를 이용한 검색 예시



출처: IBM, 2012

컨텐츠 분류

컨텐츠 분류는 컨텐츠 분석의 핵심 기능 중 하나이며, 여기에는 다양한 처리 루틴과 잠재적인 배치 전략이 포함됩니다. 일반적으로, 엔터프라이즈 검색 기능에는 고급 기계 학습 기반의 자동 분류 생성을 위한 주제 기반 클러스터링이 통합되어 있으며, 컨텐츠 분류는 컨텐츠 분석 애플리케이션 구현 작업자에게 광범위한 옵션을 제공합니다. 규칙 기반의 사전에 정의된 분류와 유연한 패싯 분류 접근법처럼 이러한 옵션 중 일부는 상호 보완적인 관계에 있습니다. 다른 옵션을 이용하려면 구현 작업자는 알고리즘의 기법과 종류 사이에서 선택을 해야 합니다.

컨텐츠 내의 주제를 분류하는 기능은 사용자가 컨텐츠 집합을 통해 성공적으로 검색할 수 있도록 하는 핵심 요소입니다. 이 기능은 드릴다운 방식의 탐색 패턴을 제공하며, 사용자는 이용 가능한 패싯을 "서핑(surf)"하여 컨텐츠의 영역을 관심 주제와 관련된 데이터 요소로 좁힐 수 있습니다 (예: 음료/와인, 국가/프랑스, 색/적포도주, 포도 품종/피노 누아, 지역/부르고뉴, 빈티지/2007년). 전자 상거래 웹 사이트에서 최초로 상용화가 시도된 이러한 패싯 탐색 접근법은 대부분의 정보 검색 및 분석 애플리케이션에서 업계 표준이 되고 있습니다.

분류는 사용자가 이용하는 기능뿐만 아니라 백오피스 또는 "lights out" 루틴의 용도로 이용될 수 있습니다(예: 아카이빙에 대해 적절한 문서만을 라우팅하고 다른 문서는 기록 관리 또는 eDiscovery 시스템의 기록의 맥락에서 검색). 컨텐츠 분석 개발 팀은 클러스터링 및 분류 소프트웨어가 제공하는 다수의 기능과 구현 전략을 숙지해야 합니다.

감성 추출

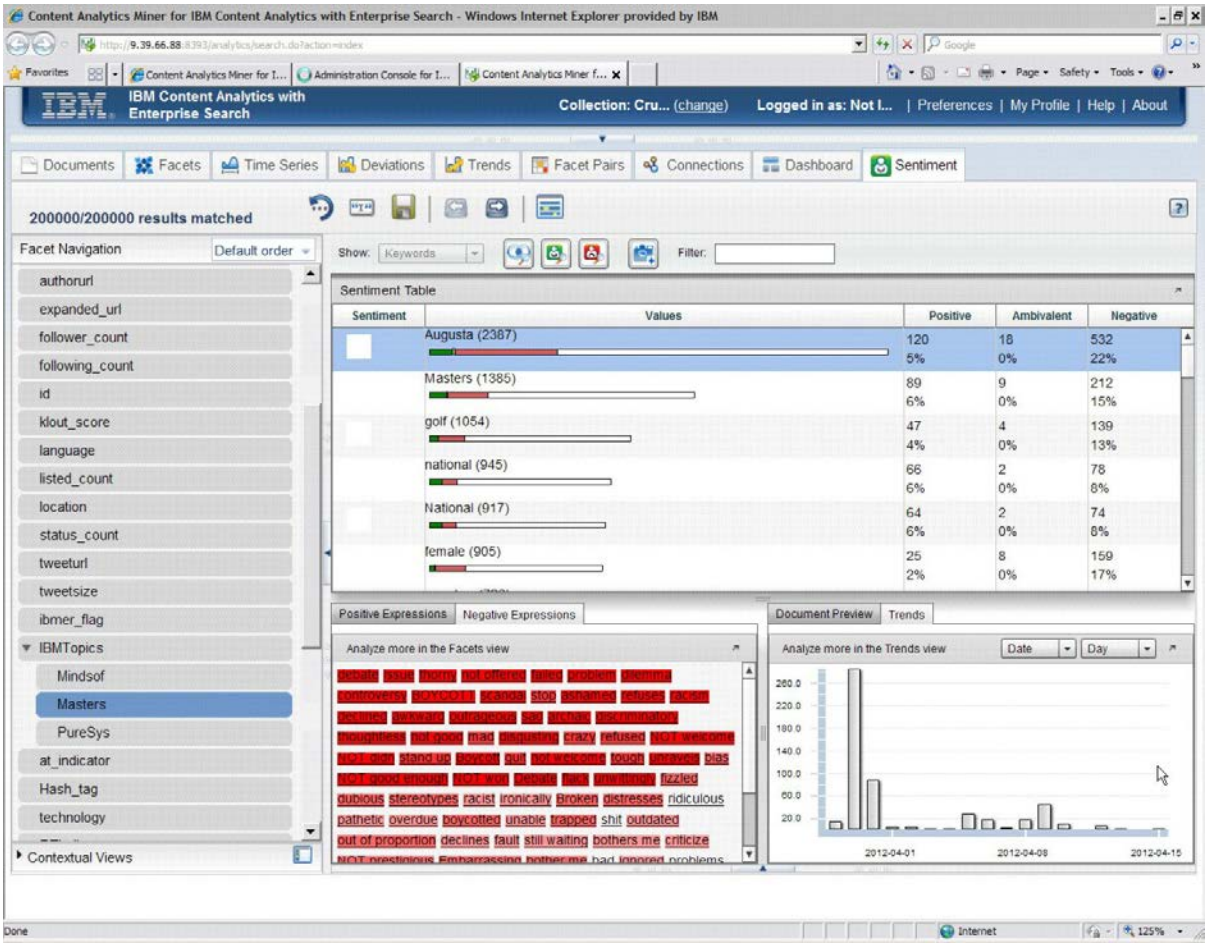
감성 추출은 컨텐츠 분류에서 특성화된 응용 분야 중 하나이며, 컨텐츠 분석 소프트웨어는 의견이나 평가를 표현하는 특정한 용어와 구절을 찾고 이를 어떠한 개체나 이벤트에 대한 긍정적인 평가, 부정적인 평가 또는 중립적인 평가로 분류합니다. 감성 추출이 상용화된 것은 최근이 아니지만, 소셜 미디어의 성장, 그리고 커뮤니티 게시물과 블로그의 사용자 생성 컨텐츠 및 기업, 제품 또는 특수 서비스를 제공하는 웹사이트(예: Yelp)에 대한 코멘트 스트림의 폭발적인 증가로 인해 감성 추출의 용도 및 가치가 크게 상승했습니다.

일상적인 대화에서 사용되는 평가와 관련된 용어는 계속해서 변화하기 때문에, 감성 추출은 분류 엔진의 활용뿐만 아니라 사전의 품질에도 크게 좌우됩니다. 가치에 대한 표현은 언어나 문화에 따라서도 달라지므로, 감성 추출은 문화에 대해서도 의존도가 높습니다. 예를 들어, 미국 보스턴 출신인 사람은 어떤 대상을 평가할 때 "wicked good(지독하게 좋다)"라고 말하는 것이 자연스럽겠지만, 미네소타 주에서는 이 표현이 어색하게 들릴 것입니다.

감성은 정보에 대한 통찰력을 제공하나, 표준 컨텐츠 분석을 이용하는 경우에는 이러한 통찰력을 얻을 수 없습니다. 감성 분석의 한 예시가 그림 4에 표시되어 있습니다. 이 예시에서는 2012년도 Masters 골프 대회에 대한 트윗이 분석됩니다.

그림 4

Masters 골프 대회에 대한 감성 분석 보고

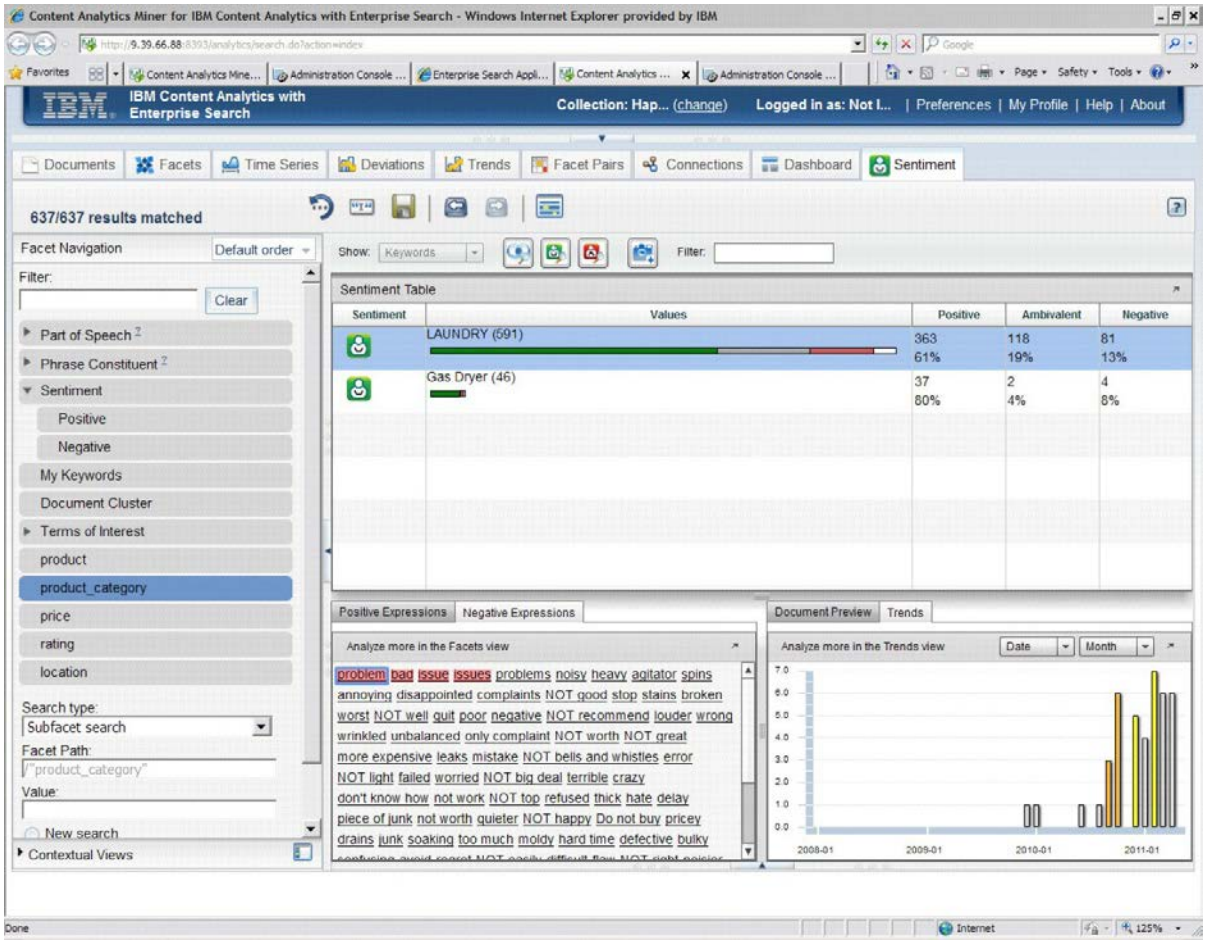


출처: IBM, 2012

감성 분석의 또 다른 예시가 그림 5에 표시되어 있으며, 여기에는 다양한 가정용품에 대한 감성이 표시되어 있습니다. 감성 분석에서 유념해야 할 것은, 감성 분석이 컨텐츠 분석 프로세스의 중요한 부분이 될 수 있다는 것입니다. 조직은 전체적인 컨텐츠 분석 전략의 일환으로 감성 분석이 필요한지 판단해야 하며, 감성 분석이 필요한 경우에는 기존 컨텐츠 분석 툴이 감성의 분석, 추출 및 표시 기능을 제공하는지 확인해야 합니다.

그림 5

가정용품에 대한 감성 분석 보고



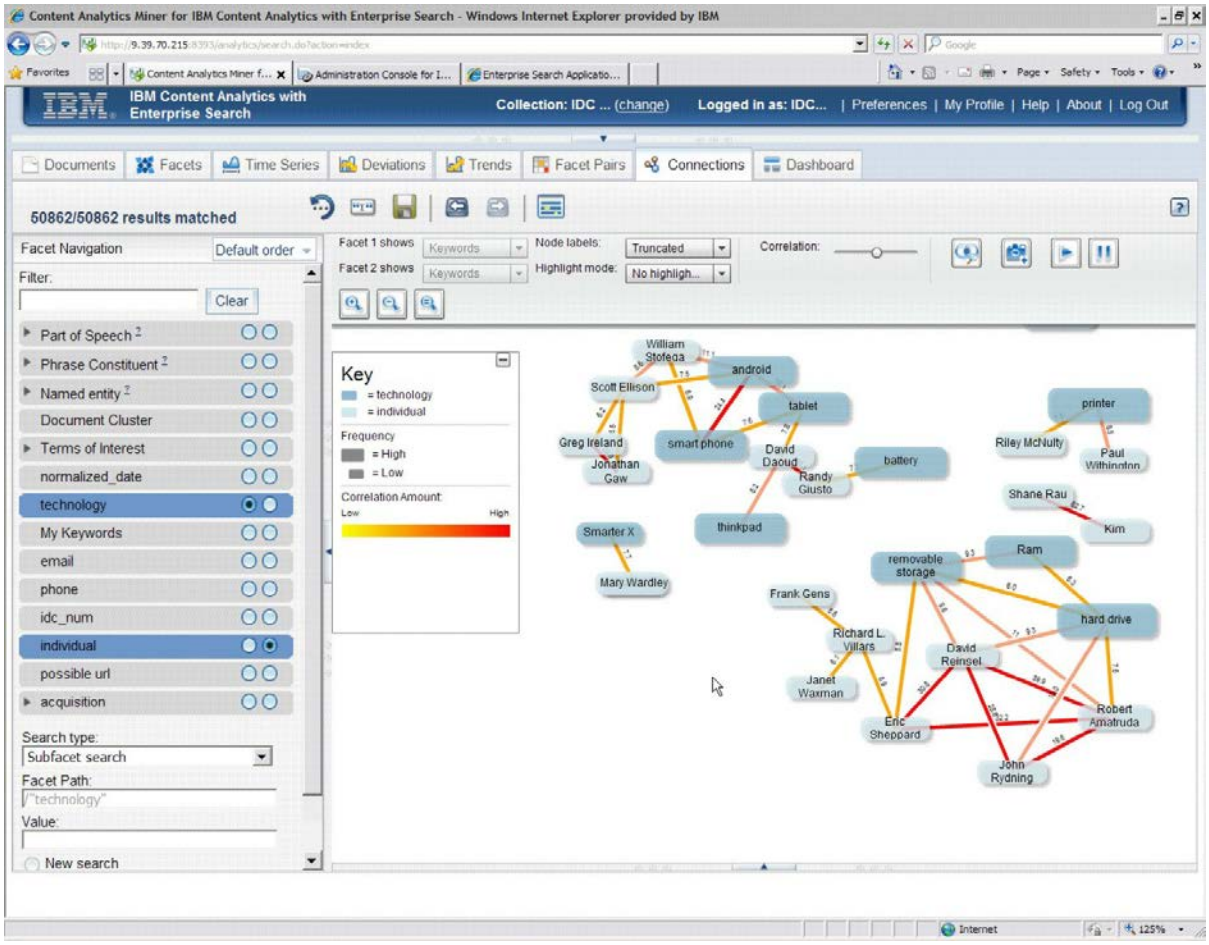
출처: IBM, 2012

시각화

시각화는 컨텐츠 분석이 제공하는 강력한 툴에 포함되어 있습니다. 시각화를 이용하면 정보와 컨텐츠를 더 쉽게 탐색하고 발견할 수 있습니다. 개체의 유형 사이의 상관관계를 통해 유형 간의 관계를 알 수 있으며 다른 유용한 정보들도 파악할 수 있습니다. 예를 들어, 그림 6에 표시된 시각화는 기사와 보고서 내에서 "기술(technology)"과 "개인(individual)"이라는 개체 항목 간의 관계를 나타냅니다.

그림 6

개체의 상관관계에 대한 도식



출처: IBM, 2012

실제 사례 소개

시각화

프로젝트의 시각화 관련 필요사항 식별

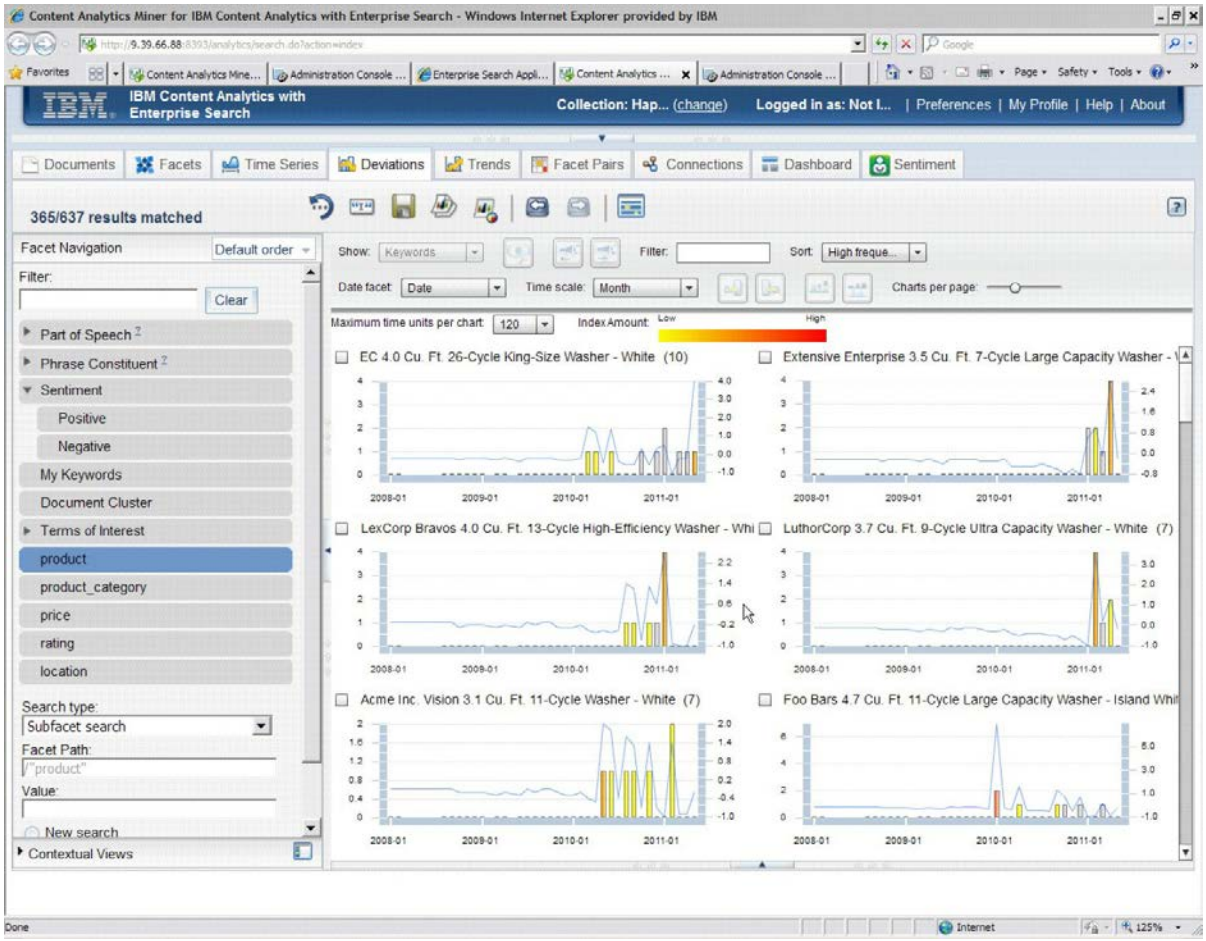
정보를 추출 또는 "마이닝"한 후, 이러한 정보를 텍스트가 아닌 차트나 그래픽으로 표시하면 사용자는 관계, 트렌드 및 패턴을 더 쉽게 이해하고 식별할 수 있습니다. 마이닝한 컨텐츠 요소에 들어 있는 데이터에 대해서는 단순한 타임라인, 막대 그래프 또는 파이 그래프를 효과적으로 사용할 수 있습니다. 데이터와 컨텐츠를 하나의 그래픽으로 결합하면 더욱 완전한 시야를 제공할 수 있습니다.

컨텐츠 분석 제품은 매우 다양하며, 제품 내에서 시각화 기능을 제공할 수 있는지의 여부에 따라 구분됩니다. 넓게 보면, 컨텐츠 분석 보고 툴과 시각화는 비즈니스 인텔리전스 제품에서 제공되는 시각화 툴과 기능적으로 유사합니다. 이러한 기능은 대규모의 맞춤화 작업 없이도 발견 및 탐색 기능을 제공할 수 있어야 합니다. 조직들은 의사결정을 위한 시각 자료에 대해 고유한 방식을 사용하고 있는 경우가 많으며, 컨텐츠 분석 소프트웨어는 이러한 그래픽 방식으로 정보를 제공할 수 있어야 합니다. 또한, 컨텐츠 분석 소프트웨어는 써드파티 시각화 툴에 대한 애플리케이션 프로그래밍 인터페이스(API) 또는 게이트웨이를 제공해야 합니다. 컨텐츠 분석 개발 팀은 프로젝트 초기에 시각화 필요사항을 거시적인 수준에서 식별한 후, 컨텐츠 분석 제품 이외의 시각화 툴이 포함되어야 할 필요가 있는지 아니면 컨텐츠 분석 제품 내의 표준 시각화 기능으로도 충분한지 판단해야 합니다.

시각화 기능에는 다양한 유형이 존재하며, 이러한 시각화 기능들은 흥미로운 방식으로 결합되어 제한된 공간에서 추가적인 정보를 제공합니다. 그림 7에서는 여러 종류의 세탁기에 대한 제품 평가가 특정 기간에 대해 비교됩니다. 이러한 유형의 시각화를 이용하면 사용자는 한 화면에서 여러 가지 개체를 직접 검토 및 분석할 수 있습니다.

그림 7

세탁기에 대한 감성 비교 보고



출처: IBM, 2012

맞춤식 분석 구성요소

대부분의 컨텐츠 분석 구현 작업에는 해당 애플리케이션 고유의 요구사항을 수행하기 위한 맞춤식 어노테이터가 필요합니다. 이러한 어노테이터의 종류는 감성 처리에 대한 고유한 접근법에서부터 통화 기록으로부터 데이터를 가져오기 위한 맞춤식 루틴, 빅데이터 클러스터로부터 소셜 미디어 활동을 마이닝하기 위한 인터페이스에 이르기까지 다양합니다.

각각의 컨텐츠 분석 소프트웨어 제품은 자체적으로 일련의 언어나 개발 마법사를 지원하여 맞춤식 모듈 또는 어노테이터를 생성할 수 있도록 합니다. 다수의 상용 소프트웨어 패키지는 새로운 산업 표준을 지원합니다. 컨텐츠 분석 소프트웨어에서 제공하는 맞춤화 기능의 범위를 평가하고 이러한 기능이 제공하는 유연성의 수준 및 사용 편의성을 측정하는 것은 컨텐츠 분석 팀에게 중요한 작업입니다. 필요한 기술적 역량과 리소스를 해당 팀이나 소프트웨어 벤더로부터, 또는 전문적인 서비스 커뮤니티로부터 제공받을 수 있는지의 여부를 평가하는 것 또한 컨텐츠 분석 팀에게 중요한 작업입니다. 컨텐츠 분석에 대한 맞춤식 개발에는 특정한 역량이 필요하며, 현재 이러한 역량에 대한 수요는 공급에 비해 훨씬 더 큽니다.

맞춤식 개발을 처리하기 위한 한 가지 방법은, 일반적인 리소스를 이용할 수 있는 경우 이를 활용하는 것입니다. 분류, 사전 및 어휘 목록에 대한 저장소와 유사하게, 대부분의 컨텐츠 분석 시스템은 맞춤식 어노테이터와 개체 어노테이터의 저장소를 갖추고 있습니다. 예를 들어, 그림 9는 IBM Content Analytics 제품에 대한 IBM의 Text Analytics Catalog를 나타냅니다. 이러한 카탈로그는 조직이 다른 조직과 함께 개발, 공유 및 이용할 수 있는 맞춤식 어노테이터로 구성되어 있습니다.

실제 사례 소개

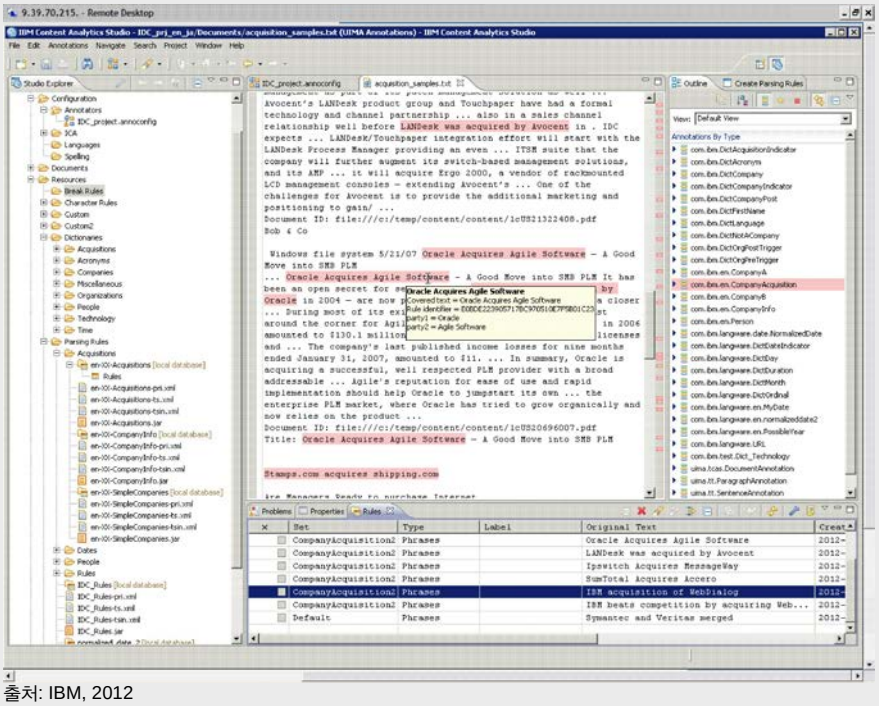
맞춤화: 스크립트와 모델의 비교

고급 모델링 툴의 활용

컨텐츠 분석 개발 팀이 소프트웨어를 평가하려면 "곧바로 이용할 수 있는" 분석 루틴이 필요합니다. 이러한 기능이 제공되지 않거나 맞춤화된 분석 루틴을 신속하게 개발하기 위해 기능을 조합할 수 없는 경우, 분석 파이프라인 내의 작업을 실행하기 위한 새로운 스크립트를 전담 개발자가 코딩해야 합니다. 이는 과거에도 수백 번 이상 개발되었을 하위 수준의 기능에 불과할 가능성이 높습니다.

최신 제품들은 다양한 범위의 곧바로 이용 가능한 루틴과 그래픽 모델링 환경을 제공하며, 이를 통해 컨텐츠 분석 파이프라인의 구성 및 맞춤화 작업의 프로세스를 간소화하여 특정한 애플리케이션 요구사항을 충족할 수 있습니다. 그림 8에 표시된 스크린 샷은 현재 고급 제품에서 제공하고 있는 인터페이스입니다.

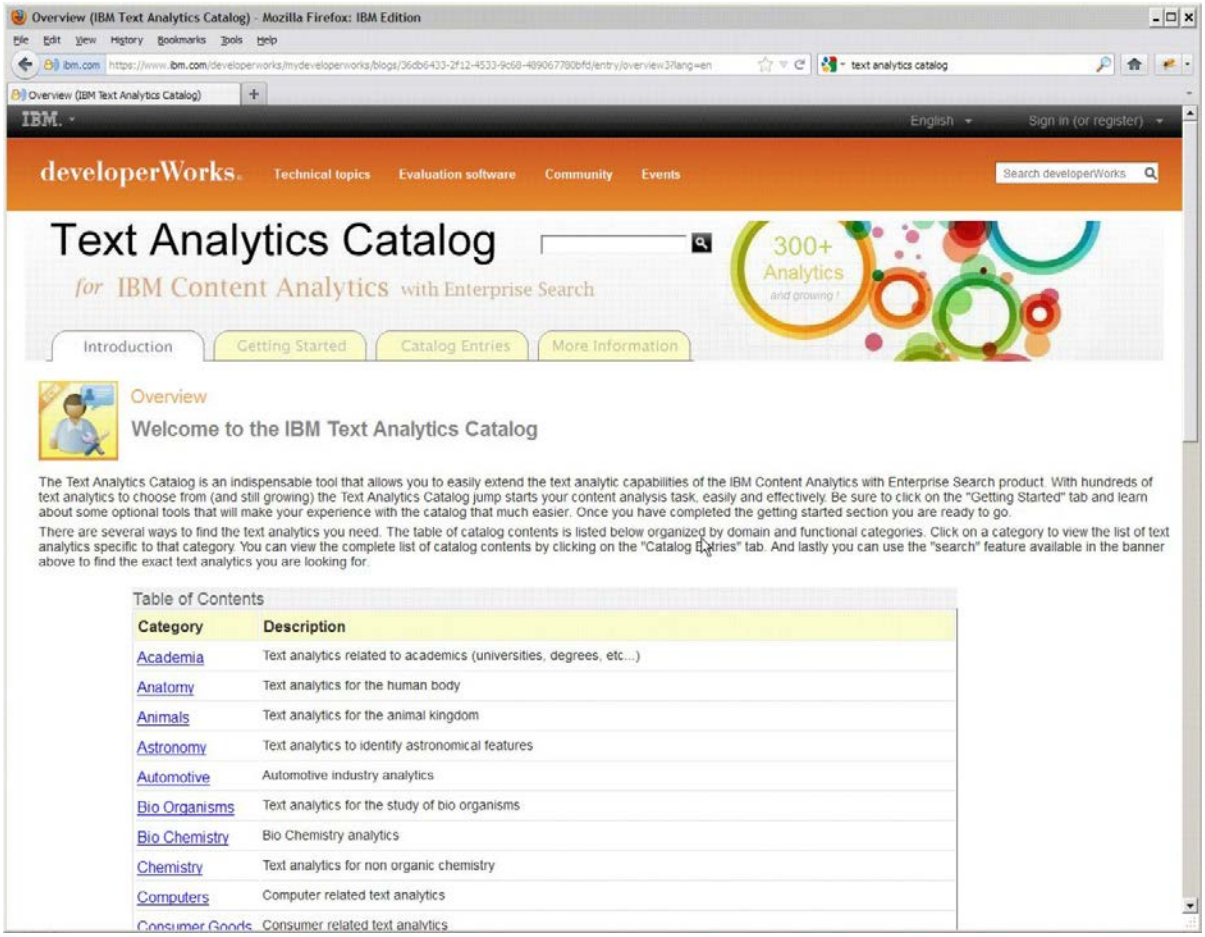
그림 8
분석 모델링 툴 스크린 샷



출처: IBM, 2012

그림 9

IBM DeveloperWorks Text Analytics Catalog



출처: IBM, 2012

과제와 기회

정보가 제공하는 이점

특정한 사실, 사람, 문서 또는 이미지를 검색하거나 정보 공간을 탐색하여 새로운 생각과 트렌드를 발견하는 정보 검색 활동은 온라인 경험에서 필수적인 부분이 되었습니다. 이러한 두 가지 프로세스 모두를 향상시키는 컨텐츠 분석은 광범위한 분석이라는 트렌드의 한 부분이기도 합니다. 조직들은 트랜잭션, 고객, 난방 및 냉방까지 모니터링하여 파악하고자 하지만, 표준 비즈니스 인텔리전스 애플리케이션만으로는 한계가 있습니다. 완전한 시야를 확보하려면 컨텐츠 분석을 통해 텍스트 및 기타 비정형 정보를 통합하여 이메일, CRM 시스템, 계약, 판매 및 수리 기록뿐만 아니라 조직 외부의 미디어 블로그, 위키 및 트윗을 마이닝해야 합니다.

기업의 직원과 소비자를 포함한 사람들의 온라인 정보 경험에 있어서 검색이 점점 중심적인 역할을 함에 따라, 이해하기 쉽고 관련성 높으며 정확한 검색 결과에 대한 요구가 발생하고 있습니다. 이에 컨텐츠 분석 기술의 도입이 점점 증가하고 있습니다. 웹 규모의 컨텐츠 분석 처리와 같은 과제를 해결해야 할 필요가 늘어남에 따라, 컨텐츠 분석 소프트웨어 공급업체의 혁신이 필요하게 되었습니다.

결론

실적상위 조직이 동종업체와 경쟁업체와 차별화되기 위한 전략에서, 분석은 가장 중요한 부분이 되었습니다. 컨텐츠 마이닝은 현재의 트렌드나 이전에 발견하지 못했던 관계에 대한 신속하고 심도 있는 통찰력을 얻을 수 있도록 합니다. 이러한 컨텐츠 마이닝을 통해 환자 치료 향상, 새로운 비즈니스 기회 포착, 시장 출시 기간의 단축, 위험의 신속한 발견과 사전 조치 등을 실현할 수 있습니다.

점점 더 많은 비즈니스 기능이 디지털화되어 가면서, 기업의 핵심 업무 중 하나인 비즈니스 분석의 중요성은 지속적으로 커지고 있습니다. 컨텐츠 분석 소프트웨어는 정형 정보와 비정형 정보라는 영역의 교차점에 위치하며, 메인스트림 분석에서의 데이터 소스 통합에서 중심적인 역할을 합니다.

기업 내의 정보의 양이 기합급수적인 속도로 증가함에 따라, 이제 컨텐츠 분석의 잠재력을 무시할 수만은 되었습니다. 정보의 홍수를 처리하기 위한 자동화된 툴은 컴플라이언스에서부터 고객 경험 및 공급망 최적화에 이르기까지, 기업의 모든 기능 영역에 걸친 광범위하고 정확한 비즈니스 분석을 수행할 수 있는 유일한 솔루션입니다.

최근 IBM 기업가치연구소의 설문조사(CEO Study, 2012년 5월, http://www-935.ibm.com/services/us/en/c-suite/ceostudy2012/)를 통해 식별된 실적상위 조직들은 다른 기업 그룹에 비해 분석을 비즈니스의 더 많은 측면에 적용하고 있으며 더욱 정교한 분석을 이용하고 있는 것으로 나타났습니다. 컨텐츠 분석은 대량의 비정형 데이터를 이용하고 비정형 데이터와 정형 데이터를 결합하여 핵심적인 비즈니스 문제를 광범위하게 처리할 수 있는 새로운 기회를 제공하고 있습니다. 이러한 문제를 신속하고 완전하게 이해하는 것은 대규모의 비용 유출을 방지하고 전체적인 시야를 확보하여 새로운 비즈니스 계획에 착수함으로써 매출을 증대시킬 수 있는 최상의 방법입니다.