'문서 데이터 웨어하우스' 가 뜬다

특성 벡터 추출 가능… '실현 가능성' 의문도 많아

전자상거래나 비즈니스 인텔리전스와 같이 데이터 웨어하우징도 현재 대부분의 기업 IT 일정표에서 막중한 부분을 차지하고 있다. 이는 신뢰성을 갖춘 시스템 유용성, 확장형 데이터베이스 서버, ETL(extract-transform-load) 툴, 온라인 분석 처리(OLAP) 애플리케이션과 같은 다양한 요소들에 힘입어 그 같은 성공을 이루게 된다.

그러나 데이터 웨어하우징의 성공을 주도한 요소는 막대한 데이터 볼륨을 손쉽게 알아볼 수 있도록 표시하는 일반화된 차원형 모델의 존재에 있다. 이러한 모델들 - 본질적인 의미에서는 스타 스키마 - 은 현재 보편화되었기 때문에 관계형 데이터베이스는 그러한 모델에 대해 최적화 되었고 다양한 에드 핫 쿼리 툴들은 스타 스키마 기반의 아키텍처를 근간으로 한다.

'차원형 모델링' 주목

다른 기술과 마찬가지로 차원형 모델링도 한계를 지니고 있다. 하지만 이 모델링의 잠재력이 아직까지도 완전히 발휘되지 못한 부분이 있다. 그것이 바로 텍스트 관리. 이 기사에서는 차원형 모델링이 데이터 웨어하우스에서 텍스트를 지원하는 방법을 알아볼 것이다.

차원형 모델이 제공하는 기술은 실질적인 비즈니스 인텔리전스의 잠재적 소스이지만 일정한 체계가 없는 방대한 양의 텍스트를 논리적이고 효과적으로 구성할 수 있게 만든다. 차원형 모델은 차원 구조에 기반한 개념을 통해 잡다한 텍스트를 일목 요연하게 재구성함으로써 텍스트의 표면적인 기능에 의존하는 기존의 간단한 기술(키워드 검색이나 문자열 일치)이 갖는 한계를 극복할 수 있게 된다.

텍스트는 종종 일정한 체계가 없는 데이터로 잘못 분류되기도 한다. 이것은 확실히 잘못된 것이다. 지난 50년 동안 변형 문법을 연구한 노암 촘스키(Noam Chomsky)와 다른 언어학자들은 언어의 기본적인 구조에서 복잡한 모델을 발견하기에 이른다.

언어가 갖는 구문과 의미에 대한 오랜 연구를 통해 마침내 이들은 텍스트 구조를 명확한 형식으로 분석하는데 도움이 되는 기초를 제공하게 된다. 따라서 실제 문서 데이터 웨어하우스에 필요한 조건들은 전통적인 숫자 기반의 웨어하우스의 필요 조건들과 유사(일부분은 동일함)하다.

문서 데이터 웨어하우스에서 문서들 - 전자메일 메시지와 같이 간단한 것부터 미 식품의약국의 새로운 약품 신청서와 같이 복잡한 것에 이르기까지 - 은 구조화된 데이터와 유사한 방식으로 데이터 웨어하우스에 수집되어야 하며, 텍스트에 담긴 정보들은 손쉬운 검색이 가능하도록 모두 색인으로 정리(아마도 하나 이상의 방법으로)되는 일정한 구조를 지녀야만 한다.

텍스트도 체계가 있어

텍스트 데이터 웨어하우스에서 스타 스키마의 사실에 해당하는

것이 문서나 텍스트 요약이다. 여기서 문서는 단어, 단어 카테고리, 혹은 제목이나 주제와 대응되는 차원 세트에 의해 확인이 가능하다.

이를테면 세계무역기구(WTO) 회의의 의사록의 경우 금융 정책, 은행, 국제무역과 같은 일반적인 주제나 태국 섬유 수입 제한과 같은 보다 구체적인 주제를 가지고 색인을 달 수 있다.

하지만 텍스트가 차원 표시에 적합하다 하더라도 관련 프로세스는 일반적인 차원 모델링과 다소 다른 면을 가진다. 따라서 문서 데이터 웨어하우스 프로세스를 다룰 때에는 전통적인 데이터 웨어하우스와 직접 비교되는 면들을 따질 것이 아니라 유사한 측면들을 고려하는 것이 좋다.

이를테면 문서 데이터 웨어하우스에서 차원은 대개 사전에 정의되는 것이 아니라 텍스트에서 추출되며, 문서 세트가 늘어날수록 차원도 똑같이 비례하여 늘어난다. 더우기 어그리게이션과 같이 정확한 일치성을 갖는 작업이 숫자 기반의 데이터 웨어하우스에서는 가능하지만 문서 테이터 웨어하우스에서는 불가능(문서 클러스터링의 경우 비록 여러 가지 방법이 제안되었지만 숫자와 같은 정확한 일치성은 불가능함)하다.

EFL과 유사

또한 동일한 클러스터링 기술 내에서도 알고리즘의 미묘한 차이 때문에 상이한 클러스터 그룹핑이 나올 수밖에 없다.

텍스트 데이터 웨어하우스는 ETL과 비슷한 면이 많다. 문서에 대한 ETL 프로세스의 주요 단계는 메타데이터 기준에 기초한 필터링, 사람이나 조직 이름과 같은 특성 추출, 키워드와 주제에 대한 색인 달기, 문서 요약, 관련 클러스터로 문서 모으기(그룹핑) 순서로 이루어진다.

필터링. 텍스트 데이터는 숫자 데이터와 달리 구조적 한계를 가지고 있지 않으므로 정확하지 못한 값을 확인하고 완벽히 제거하는 것은 일단 제외 대상이다. 문서 데이터 웨어하우스의 경우 제거에 해당하는 비슷한 것이 메타데이터 정보나 내용에 기초한 필터링으로서 개별 소스에 대한 메타데이터가 필터링의 첫 번째 기준이 된다.

이를테면 변호사나 대리인이 작성하는 문서를 제외시킴으로써 보호되는 기밀 문서의 보관을 위해 당신은 판촉활동 문서와 메일 메시지를 한 묶음으로 유지할 수 있다. 메시지 내용은 필터링의 두 번째 기준으로서 예를 들어 특정 파티와 주고 받는 메일 메시지, 메일링 리스트를 통한 메일 메시지, 특정 단어나 구를 포함하는 메시지를 차단하는 경우를 들 수 있겠다.

또한 문서 종류와 소스에 대한 메타데이터는 문서에서 수행되는 작업을 결정짓는 요소로서, 이를테면 메타데이터는 웨어하우스가 문서 요약과 URL이 갱신될 때 전체 문서를 저장할지, 아니면 문서의 요약이나 URL만을 저장할지를 결정할 수 있으며, 키워드 색인 작업이 충분한지, 혹은 주제 색인 작업이 필요한지를 선택할 수도 있다.

한편 특정 알고리즘을 통한 적절한 클러스터링이 제대로 안되는 이유는 일상회화에서 사용되는 구어체에 기인한다. 따라서 개별 데이터 소스에 대한 클러스터링 방법을 확인할 필요가 있다.

다음 단계인 특성 추출 과정에서는 문서 클러스터링에 대한 기초를 제공하는 단어나 구와 같은 중요한 어휘 아이템을 통계적으로 확인하는 절차가 수반된다. 기본적인 특성에는 사람과 조직의 이름, 온라인 분석 프로세싱과 같은 복합명사, 약어, 그리고 알파 Analysis나 베타 소프트웨어 시스템과 같은 고유 명사가 포함된다.

색인 달기. 텍스트 검색을 위한 색인 달기는 보통 키워드와 주

제를 이용한다. 키워드 검색은 문서 내에 포함된 특정 단어의 위치를 추적하는 방식으로서 수동적으로 키워드를 확인하는 방법과 자동으로 키워드 리스트를 생성하는 방법이 있다.

이 자동 리스트는 문서에서 자주 사용되는 단어를 검토하여 가장 많이 사용되는 단어(이러한 단어들은 문서상에 가장 많이 분포되어 있기 때문에 문서 구분에 별 도움이 되지 않음)를 제외시킴으로써 생성될 수 있다.

자동 리스트가 아니더라도 일반적으로 많이 사용되는 단어에 대한 사전 정의된 리스트를 이용함으로써 정보를 담지 않거나 정보가 거의 포함되지 않은 단어를 가려낼 수 있다. 이러한 방법은 일반적인 문서 검색에 유용하며, 자동 생성 리스트의 경우는 특정 문서 검색에 적합하다.

한편 주제 색인 방식은 키워드 색인에 비해 보다 포괄적인 검색 기능을 제공한다. 키워드 검색의 경우 사용자는 아스피린과 마취제와 같은 유사한 2개의 단어를 모두 검색해야 하지만, 주제 검색에서는 진통제라는 카테고리에서 한 번에 찾을 수 있다.

주제 색인 방식에서는 키워드 방식과 마찬가지로 사전 정의 방법을 사용할 수 있으며, 또는 문서 집합을 바탕으로 한 분류 방법을 사용할 수 있다. 후자의 경우 IBM 인텔리전트 마이너 포 텍스트 카테고리제이션 툴(Intelligent Miner for Text categorization tool)을 구성하는 여러 텍스트 분석 도구중 카테고리화 도구를 이용할 수 있다.

요약. 숫자 데이터 웨어하우스는 고도의 계산 평가를 위해 어그리게이션에 의존하며, 문서 데이터 웨어하우스는 불필요한 세부 정보를 제거하고 핵심 정보를 제공하기 위해 문서 요약을 사용한다. 요약 크기는 문서 크기와 문서가 담고 있는 정보의 양에 따라 달라진다.

요약. 크기가 클수록 포함된 정보는 많겠지만 실질적인 크기 제한선이 존재한다. 정보의 양과 요약 크기를 어느 한쪽에 치우치지 않고 적절히 조화시키려면 필연적으로 시행착오를 거쳐야 할 것이다. 전자메일, 메모, 정책 문서와 같은 종류의 문서들은 원래 텍스트 크기에 비해 비교적 작은 크기로 요약하기에 적당하다.

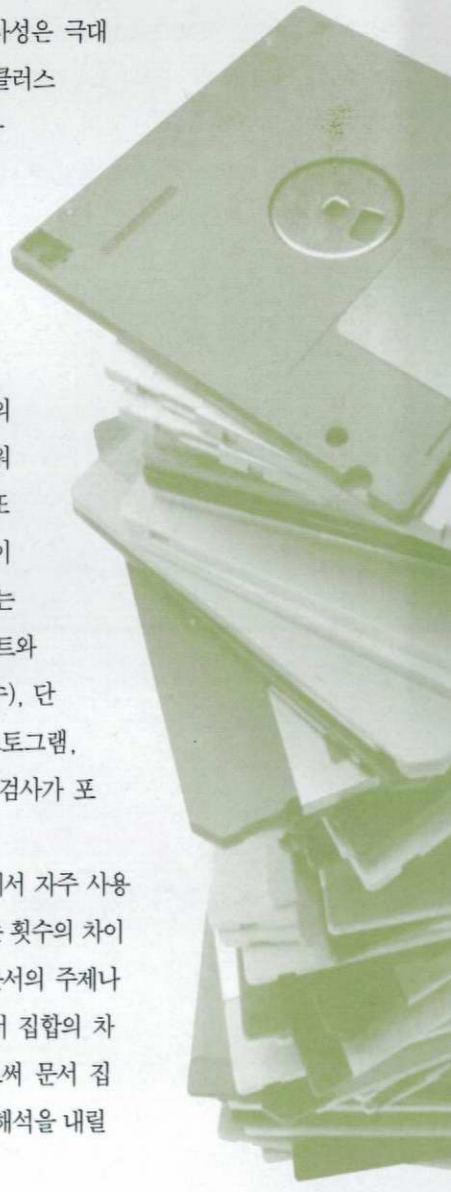
그러나 의학이나 법률과 같은 전문 분야의 문서들은 적절한 정보 내용을 유지하기 위해 보다 세부적인 정보를 포함하는 요약이 필요하다. 일반적으로 문서 종류에 대한 메타데이터는 요약 정도를 제어해야만 한다.

어차피 요약이란 문서 자체를 의미하기 때문에 동일 기술이 특성 추출, 색인 달기, 클러스터링과 같은 다른 단계에도 사용될 수 있다. 그렇지만 항상 좋은 결과를 가져오는 것은 아니다.

클러스터링. 대부분의 웹 검색 엔진은 키워드 검색 결과에 '동일한 문서를 찾을까요' 라는 옵션을 같이 제공한다. 합계와 같은 계산 작업과는 달리 텍스트 검색에서는 유사한 것에 대한 고정된 공식 정의는 존재하지 않는다.

따라서 문서의 클러스터링이나 그룹핑과 관련된 문제에 있어 완벽한 단일 솔루션은 존재하지 않는다. 그럼에도 불구하고 몇 가지 방법이 꾸준히 개발되었는데, 대략 세 가지로 나눌 수 있다: 계층형 클러스터링, 바이너리 관계형 클러스터링, SOM(self-organizing map) 클러스터링.

세 가지 방법 모두 인트라클러스터(intracluster)의 유사성은 극대화하는 동시에 인터클러스터(intercluster)의 유사성은 최소화하는 것을 목표로 한다.



클러스터링에는 두 가지 조건이 필요한데, 하나는 텍스트의 특성(상호 비교가 쉬워야 함) 표시이며, 또 하나는 유사성 측정이다. 일반적인 표시에는 특성 벡터(단어 리스트와 리스트가 발생한 횟수), 단어 카테고리의 히스토그램, 문서 주제의 가중치 검사가 포함된다.

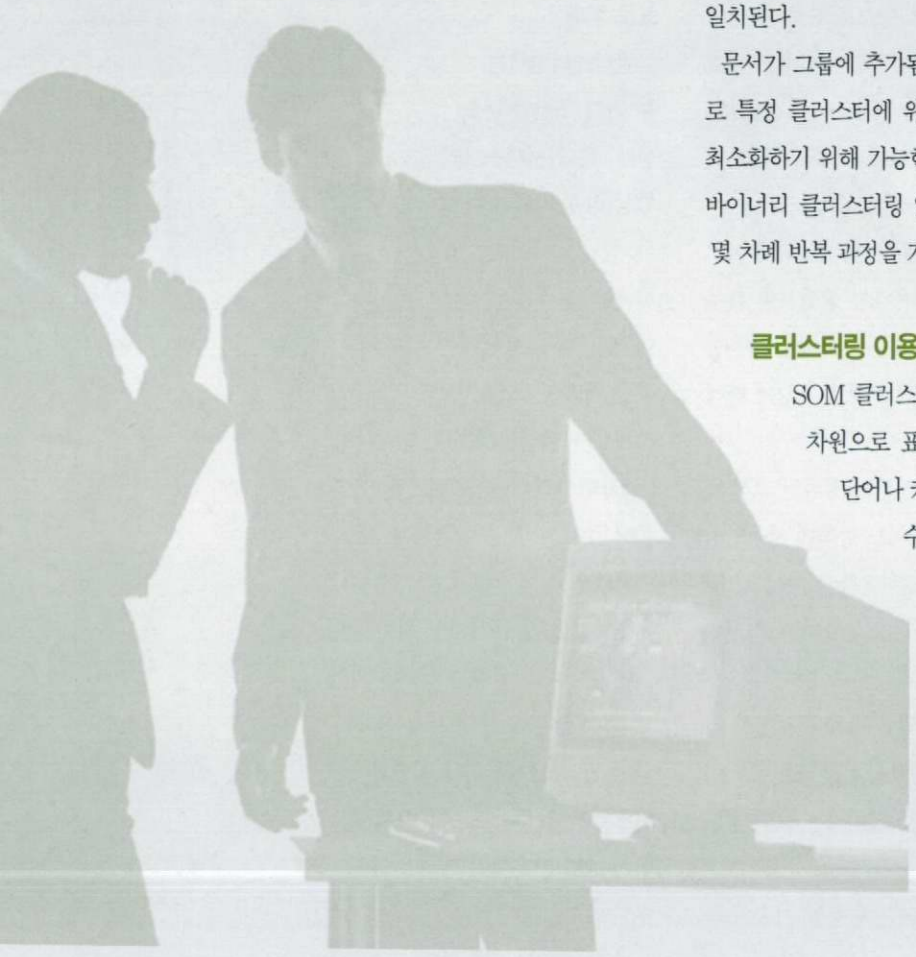
유사성은 각 문서에서 자주 사용되는 단어가 발생하는 횟수의 차이로 측정된다. 또한 문서의 주제나 단어 카테고리를 문서 집합의 차원으로서 간주함으로써 문서 집합에 대한 기하학적 해석을 내릴 수도 있다.

"유사성을 찾아라"

기하학적 해석의 경우 문서내 발생 횟수나 주제의 가중치는 문서의 차원을 측정하는 기준이 된다. 차원 세트에 의해 정의되는 다차원 공간상의 문서 위치는 특성 벡터[d,, d2, d3, dn]에 의해 알 수 있는데, 여기서 dj는 차원상의 주제 가중치나 발생 횟수를 말한다.

이러한 기하학적 해석을 통해 계산된 문서들간 유클리드의 기하학적 거리를 모두 최소화함으로써 유명한 K-NN 데이터 마이닝 알고리즘과 똑 같은 아웃풋을 도출하게 된다. 문서 하나가 보통 10개 정도의 주제를 포함한다고 볼 때 문서 집합은 수천 개 이상의 주제를 가지게 될 터이고, 따라서 수천 개 이상의 차원을 가지게 된다.

하지만 다행히도 튜보 코호겐(Teuvo Kohonen)의 최근 연구 결과에 의하면 수천 개 이상의 고차원 표시를 수백 개 정도의 저차원 표시로 무작위 투영함으로써 모든 차원 수를 최소화할 수 있으며, 따라서 계산양을 상당히 줄일 수 있다고 한다.



당연히 식별력의 아무런 희생 없이도 가능하다는 이야기이다(인공 신경망에 대한 제 8회 국제 컨퍼런스 소식지 1998년 봄호, Teuvo Kohonen의 Self-Organization of Very Large Document Collections: State of the Art. 참조).

계층형 클러스터링은 트리 형태를 가지고 있는데, 루트는 전체 문서 집합을 의미하며 잎 노드는 개별 문서를 갖고 있는 독립된 집합을 의미한다. 계층형 클러스터링은 가지 노드와 잎 노드와 매우 유사한 노드를 하나로 결합시킨다.

따라서 가지 노드의 특성은 구성된 노드의 결합된 특성을 의미한다. 결합 과정이 계속됨에 따라 단 하나의 노드인 루트만이 남을 때까지 클러스터의 새로운 레벨이 생성된다. 계층형 클러스터링의 장점은 일반적인 그룹핑에서 부터 보다 구체적인 클러스터에 이르기까지 그 사이를 자유로이 이동할 수 있는 데이터 구조를 생성할 수 있다는 점이다.

반면 바이너리 관계형 클러스터링은 수평적 데이터 구조를 가진다. 수평적 구조에서 개별 문서는 자신을 가장 잘 나타낼 수 있는 클러스터에 위치하며, 그 결과 개별 클러스터는 특정 주제와 일치된다.

문서가 그룹에 추가됨에 따라 결합된 특성 벡터는 변하게 되므로 특정 클러스터에 위치한 문서들은 인터클러스터의 유사성을 최소화하기 위해 가능한 경우에 한해 다른 클러스터로 이동한다. 바이너리 클러스터링 알고리즘은 최상의 문서 위치를 찾기위해 몇 차례 반복 과정을 거치게 된다.

클러스터링 이용

SOM 클러스터링 기술은 산재한 고차원의 데이터를 2차원으로 표시한다. 문서 클러스터링의 경우 차원은 단어나 카테고리 중 하나이며 차원 측정은 발생 횟수와 관련되어 있다. SOM은 신경망으로 표시되며, 이 신경망의 개별 노드는 가중 벡터를 포함한다. 개별 문서는 특성 백터로서 망에 표시되며 개별 노드에 대한 거리는 유클리드의 기하학적 거리인 dk, w - dk(t) =||xt) wk(t)|| 2-로 계산된다.

가장 적합한 노드는 dk(t)의 최소값이며 그 노드에 대한 가중 백터는

다음과 같은 식에 따라 조절되어 진다.

$wk(t+1) = wk(t) + a(t)hck(t)\,[x(t) - wk(t)]$

a(t)는 비율 인자이며 hck(t)는 가중 백터의 조절양을 제어하는 이웃 함수를 의미한다. 결과적으로 망의 개별 노드는 노드와 결합된 문서 집합을 가장 잘 설명하는 가중된 특성 백터를 가지게 된다.

SOM은 WebSOM 프로젝트의 인터넷 문서 클러스터링에 사용된 적이 있으며 구어체를 포함하는 문서의 클러스터링에도 적합하다.

문서 데이터 웨어하우징의 최종 목표는 숫자와 텍스트 정보를 단일 저장소 내에서 통합하는 것이므로 저장과 검색 툴은 반드시 숫자와 텍스트 데이터를 모두 지원해야만 한다. 또한 관계형 데이터베이스상에서 용량이 큰 바이너리 객체의 사용은 문서 집합을 저장하고 관리할 수 있는 메커니즘을 제공한다.

문서 검색이 시발점

하지만 SQL은 기본적인 문서 검색에 대한 적절한 지원을 위해 확장 기능을 필요로 하는데, 현재 이러한 기능을 갖춘 몇 가지 제품이 이미 시장에 출시되었다. 일례로 오라클 콘텍스트 (지금은 오라클 인터미디어의 구성 부분)의 출시로 인해 오라클 RDBMS는 퍼지 검색, 어간 검색, 근접 검색, 가중, 누적 기능을 위한 텍스트 연산자를 지원하고 있다.

이러한 텍스트 연산자를 통해 주제 색인, 문서 요약, 문서 클러스터링이 가능해짐에 따라 이제 문서 데이터 웨어하우징은 기업 비즈니스 인텔리전스의 확장에 생명력을 불어넣고 있다.

이제 문서 데이터 웨어하우스의 시대가 도래하였다. 문서는 비즈니스 인텔리전스 운영에 있어 잠재적인 풍부한 리소스였지만, 오랜 시간 동안, 아니 최근까지도 자신의 잠재력을 드러내지 못하고 잠들어 있던 상태였다.

현재 텍스트의 전략적 가치 이용을 가능케 하는 설계 기술과 관련 툴이 개발자를 지원하기 위해 자기 자리를 지키고 있으며 데이터 웨어하우징을 통해 얻어진 다양한 경험들이 문서 데이터 웨어하우징에 적용되고 있다.

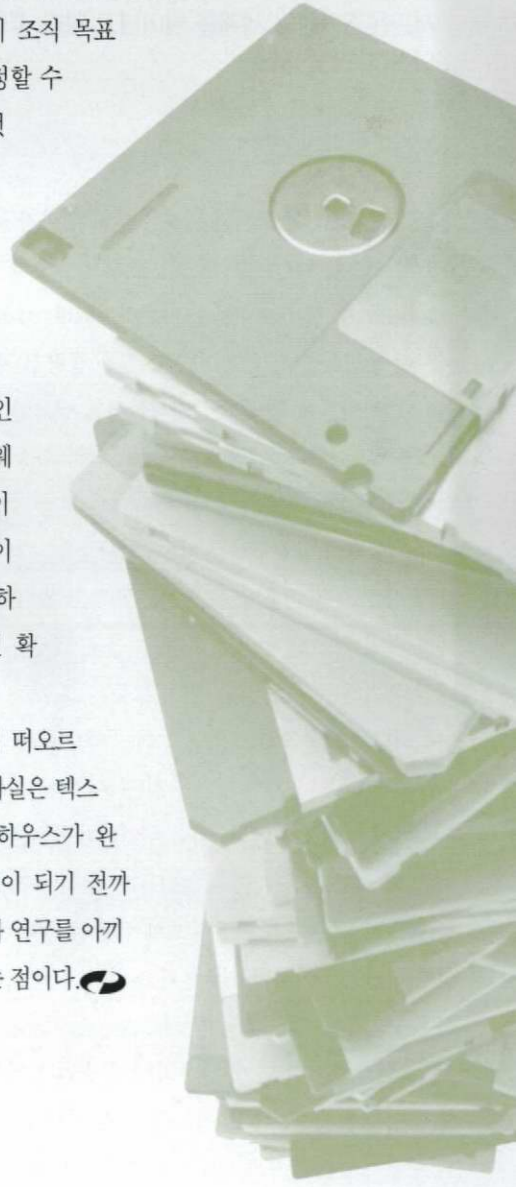
그러나 문서 데이터 웨어하우스에 관한 모든 질문들이 전부 답변을 얻은 것은 아니다. 텍스트에 대한 차원형 모델과 숫자 모델 간의 적절한 통합 방법은 무엇인가? 숫자 데이터에 사용된 동일한 차원을 가지고 문서에 색인을 달 때 가장 적합한 기술은 무엇인가?

그리고 생성 할당 요소에 대한 스타 스키마의 경우, 사용자는 아웃풋이 기대치보다 저조한 이유를 알아내기 위해 생성 제어 문서를 검토할 수 있어야 한다.

문서 웨어하우스 시대 도래

또한 복잡한 데이터는 복잡한 보안을 필요로 하는데, 이것만으로도 각종 의문이 꼬리를 물기 시작한다. 이를테면 개별 문서에 대한 액세스가 제한되는 동안에 일반적으로 액세스 가능한 문서들의 요약 범위를 검토함으로써 침입자가 모을 수 있는 정보는 어떤 것인지 궁금해질 수 있다.

이 외에도 생각나는 질문을 적어보면 다음과 같다. 특정 문서가 데이터 웨어하우스에 가장 먼저 들어가는 것은 어떤 기준에 근거한 것인가? 정보 내용이 조직 목표와 관련하여 측정할 수 있는 방법은 무엇인가? 텍스트를 위해 수집해야 하는 상이한 메타데이터는 어떤 것들인가? 일반적인 웨어하우스 모델이 이러한 메타데이터 종류를 지원하는가? 그렇다면 확장은 가능한가?



많은 질문들이 떠오르겠지만 중요한 사실은 텍스트 데이터 웨어하우스가 완전한 메인스트림이 되기 전까지 꾸준한 관심과 연구를 아끼지 말아야 한다는 점이다.