시소러스와 토픽맵의 연관성 연구*

A Study on the Association between Thesaurus and Topic Map

남영준 중앙대학교 문헌정보학과 교수

Nam, Young-joon

Professor, Dept. of LIS, Chungang University

<초록>

현재 정보검색분야에서는 검색도구로써 시소러스가 갖는 장점에도 불구하고 기존에 개발된 시소러스의 유지관리와 활용이 극히 제한적으로 이루어지고 있기 때문이다. 왜냐하면 정보의 급격한 증가로 인하여 전통적인 시소러스의 구조와 유지관리, 활용기법으로는 현대 정보의 홍수 현상에 적극적으로 대처하는데 한계에 직면하였기 때문이다. 이러한 한계점을 극복하기 위해 토픽맵의 구축알고리즘이 절대적으로 필요하였다. 이에 따라 본 연구에서는 토픽맵의 기본요소인 토픽과 대상물, 연관관계, 토픽타입 등을 이용한 시소러스 구조화 알고리즘을 제안하였다. 특히 토픽맵의 기본 요소가운데 대상물(occurrence)은 시소러스의 검색효율가운데 정도율의 확보를 가능하게 하며, 시소러스의 구축에 필요한 지식베이스의 역할을 수행하는 주요한 기법임을 확인하였다.

주제어: 토픽맵, 지식관리, 시소러스, 온톨로지

1. 서론

검색효율을 높이기 위한 노력은 과거로부터 지금까지 정보학 분야에서 추구하는 일관된 목표가운데 하나이다. 특히 정보가 무한에 가까운 규모로 증가하는 인터넷시대 검색효율은 21세기 정보화 시대가 지향하는 목표이다.

한편 오늘날의 정보의 종류는 전통적인 자료에 대한 문서파일형태를 비롯하여 웹문서, 전자메일 등과 같은 방대한 비정형 정보가 혼재되어 있는 상황이다. 특히 인터넷상에 존재하는 비정형 데이터들은 도서관의 자료관리시스템에 비해 매우 부정확하고, 비효율적이며, 제한적인 방법으로 관리되고 있다. 이에 따라 검색엔진을 이용한 전자자원관련 검색은 실질적으로 정보의 과부하를 가중시키기 때문에 오히려 검색효율이 저하되고 있다. 왜냐하면 대부분의 검색엔진은 해당 자료에 출현한 키워드의 출현빈도에 근거하여 검색이 이루어지기 때문에 이용자가 입력한 검색어가 내포한 의미를 적절하게 표현하지 못하기 때문이다. 즉, 웹이라는 색인되지 않은 방대한 데이터베이스에 대한 단순한 키워드 검색기법이 갖는 한계이다.

문헌정보학 분야에서는 이러한 문제점을 극복하고 대용량의 인터넷 정보시대의 안정적인 검색효율을 확보하기 위해 많은 연구를 경주하고 있다. 도서관은 전통적으로 주제명 표목표와 분류표와 같은 색인도구를 이용하여 효과적인 정보검색을 시도하였다. 전통적인 검색도구는 적정한 규모의 데이터베이스 검색에는 효과적이나 대용량 규모의 데이터베이스를 대상으로 하는 검색과정에서는 적정한 검색효율을 제공하지 못한 제한점을 갖고 있다.

* 본 연구는 한국과학재단 기초연구프로그램(R01-2003-000-11588-0)의 지원으로 연구되었음.

이러한 제한점을 극복하기 위해 도서관은 기존의 검색도구를 의미적으로 구조화하였다. 시소러스는 주제명 표목표나 분류체계가 갖고 있는 평면적인 키워드 배열을 특정 용어에 대한 개념적 구조화를 표현한 진일보한 검색도구이다. 시소러스는 주제명 표목을 분류체계가 갖는 계층성과 연관성을 이용하여 전체 표목(디스크립터)을 개념적으로 구조화함으로써 기존 검색도구가 갖는 키워드 매칭과 포괄적 단순검색의 단점을 극복하다.

시소러스는 기본적으로 인쇄자료의 효율적 검색을 의도한 검색도구이기 때문에 인터넷 자원 검색을 위해서는 인터넷자원의 특성을 고려할 필요성이 제기되었다. 예를 들면, XML로 표현되는 인터넷 자원의 구조적 정보와 전문검색에 필요한 언어적 및 의미적 표현을 수용할 수 있는 시소러스로 변화되어야 한다. 특히 대부분의 시소러스는 특정 주제 분야의 마이크로적 성격을 갖고 있기 때문에 디스크립터가 갖는 의미적 유한성과 다른 개념과의 연관성을 유지할 경우에 검색효율의 개선 효과를 얻을 수 있다.

XTM(XML Topic Map)은 토픽과 연관관계, 대상물을 이용하여 개념과 실물 정보, 타 정보와의 관계를 표현하는 토픽맵의 표준 온톨로지로써 특정 주제 영역의 용어간 개념을 구조화하여 의미적 관계성을 표현할 수 있다. 일반 시소러스의 디스크립터는 특정 영역의 자료를 검색할 수 있는 후조합 색인언어의 특성을 갖고 있으나, 토픽 맵은 색인된 특정 용어에 배정된 정보리소스도 검색하는 종합적인 검색도구이다. 따라서 토픽 맵은 모든 분야를 수용하는 것보다 특정 분야의 개념 구조화에 상대적으로 유리하다.

본 연구에서는 이 점에 착안하여 시소러스의 디스크립터 구조화 과정을 토픽 맵의 구축 알고리즘을 이용하여 특정 영역의 마이크로 시소러스 구축 알고리즘을 제안하고자 한다.

2. 토픽맵의 관련 연구

토픽맵은 2000년에 정식으로 세계 표준안이 만들어졌으며, 2002년 5월 그 개정판이 발표되었다.(ISO/IEC 2002) 토픽맵은 온톨로지기반의 색인어 지도로써 컨텐츠 관리까지 가능하도록 하는 개념이다. 초기 연구는 외국에서 주로 이루어졌으며, 토픽맵의 표현에 관한 것과 이를 이용한 검색 적용으로 구분할 수 있다. 특히 토픽의 표현을 그래픽 형태의 트리 구조를 적용하여 검색효율을 높일 수 있다는 연구(Benedicte Le Grand, Michel Soto 2000)가 있다. 정준원(2003)은 그래프 기반탐색은 정보의 연관 관계와 구조를 잘 표현할 수 있으나, 수많은 노드로 이루어진 Topic과 간선으로 인해 오히려 정보를 파악하기 어렵다는 부정적인 의견을 제시하고 있다. 이와 같은 지적은 시소러스의 표현형식가운데 국제 도로연구센터(International road research documentation) 시소러스의 화살표 표시 방식과 유사한 표현방식이 사용자의 가시성을 떨어트려 오히려 효율성을 저하시키는 것과 동일한 단점을 갖는 것이다.(남영준 2002)

국내에서의 토픽맵 연구는 활용적인 측면과 기술적인 측면의 연구로 구분할 수 있다. 활용과 관련된 연구는 e-커머스분야의 활용이 상대적으로 많았다. 정원규(2003)는 동서양의 철학사상가운데 벤담의 사상을 토픽맵 개념으로 분류 및 연관관계를 토픽으로 처리하였다. 그는 토픽맵이 갖고 있는 하나의 의미를 갖기 위해서는 컴퓨터가 직접 용어의 맥락과 의미를 이해할 수 없으므로 필요한 자료에 사람, 특히 전문가가 일일이 분류기호(tag)를 기입해 주어야 한다는 한계를 인정하였다. 고세영(2003)은 이 기종간의 상품분류체계를 통합하기 위한 도구로써 토픽맵을 이용하였다. 그는 이 연구에서 분류체계의 계층관계와 연관관계를 정의하고 이를 모델링하고 이를 통합하는 방법을 사용하였다. 고유미(2005)는 특허분야의 이종간 분류체계를 통합하는 방안을 제시하였다. 특허문서의 서지정보를 토픽으로 설정하여 정보서비스 목적에 따라 온톨로지를 모델링하는 방법을 사용하였다.

한편 기술적인 연구로써 정호영 등(2003)은

토픽맵의 XTM을 이용하여 부가적인 메타 데이터-토픽, 어커런스, 토픽과 토픽간의 연관관계-를 기술함으로써 키워드 검색뿐 아니라 세미나 자료 지식에 대한 생성/유지/관리의 용이함을 지원하며 토픽들 간의 분류/연관관계에 의한 검색을 지원하는 방안을 제시하였다. 이은아(2003)는 시맨틱 네비게이션 시스템은 토픽맵을 그래프로 브라우징할 수 있는 실험적인 네비게이션 시스템을 구축하여 이용자 중심의 토픽맵을 구성하고 추출하여 토픽을 효율적으로 보여줄 수 있는 토픽맵 어플리케이션을 개발하였다. 정준원(2003)은 지식맵의 효율적인서비스 방안을 제안하고자 지식맵의 일종인 TopicMap을 이용한 탐색 및 캐쉬를 이용한 정보전송 기법에 대해서 제안하였다. 그는 토픽맵을 연관성 중심으로 탐색을 지원하는 환경을 제안하고, 이 환경에서 연관된 정보의 캐쉬를 생성하여 전송하는 방법을 사용하였다. 유우종 등(2004)은 워드넷을 하나의 온톨로지로 적용하여 워드넷에 수록된 단어를 토픽으로 처리하고, 단어간 연결은 연결망으로 적용할 수 있는 방안을 제시하였다.

3. 토픽맵 모델

토픽맵은 ISO와 함께 IEC(the International Electrotechnical Commission, 국제전자기술협의회)에서 공동으로 개발한 국제 표준이다. 본 표준에서는 토픽맵을 다음과 같이 기본요소와 구조로 구분하여 정의하고 있다.

3.1 토픽맵의 기본요소

토픽맵은 특정한 개념을 나타내는 토픽의 개념과 연관성을 다른 개념구조간의 상호교환을 위해 정의한 일련의 표준화된 체계이다.(ISO/IEC 2002) 즉 하나이상의 상호연관성을 갖는 문서의 토픽맵이라 한다. 일반적으로 토픽맵은 다음과 같은 기본적인 구성요소를 갖는다.

- 대상물(occurrence) : 토픽에 해당되는 정보원의 주소
- 연관관계(association) : 토픽간의 관계 표시

이를 구체적으로 설명하면 토픽맵은 기본적으로 토픽을 비롯하여 연관관계, 대상자료 등 3개로 구성된다. 토픽은 원칙적으로는 문자 그대로 대상문헌의 핵심 개념들이 해당된다. 예를 들면, 토픽은 유형의 모든 것과 인간이 생각할 수 있는 무형적인 것까지 단어로 표현될 수 있는 것이다. 연관관계는 특정 토픽간에 의미있는 관계가 성립할 경우, 해당 관계를 의미한다. 내용면에서는 토픽과 토픽의 연결이지만 형식적으로는 토픽과 크게 구분될 바 없다. 마지막으로 대상물은 토픽이나 연관관계에 연결되는 자료를 의미한다. 여당과 야당이라는 개념을 토픽맵으로 설명하고자 한다면 토픽은 '여당', '야당', 대상물은 {여당의원1, 여당의원2, 여당의원3 … }, {야당의원1, 야당의원2, 야당의원3… }, 연관관계는 '정당 반대관계' 등으로 설명될 수 있다.

토픽맵은 하나의 다차원적 토픽 공간으로 정의된다. 여기에서 공간(space)은 토픽의 모여있는 하나의 개념사전이며, 특정 토픽간에 존재하는 토픽들이 일목요연하게 정렬되어 있어야만 하고, 하나의 토픽과 다른 토픽간의 경로가 정의되어 있는 관계가 설정되어 있는 것을 의미한다. 예를 들면 두개의 토픽은 이 공간에서 하나의 연관관계로 연결되어질 수 있으며, 또한 하나의 대상물로써 연결되어질 수 있다.

또한 정보 객체는 특성(property)과 내부적으로 부여된 해당 특성의 값을 가질 수 있다. 이들 특성을 패싯형태(facet types)로 표현한다. 예를 들면, 특정단어의 패싯은 다면체의 한면 혹은 가공된(polished) 결과물, 혼합관점에서 한 측면이라 정의할 수 있다. 은유적인 표현으로써 특성은 특정 사상을 인식할 때 사용되며, 하나의 패싯은 새로운 관점을 생산하는데 사용되는 정보객체의 특성이 될 수 있다.

여러개의 토픽맵은 동일정보원에 대해 토픽구조정보를 제공할 수 있다. 토픽맵의 구조는 다른 토픽맵을 복사하거나 수정할 필요없이 토픽맵간 융합(merging)을 할 수 있다. 왜

나하면, 비종속적인 특성 때문에, 토픽맵은 일련의 정보객체에 그대로 오버레이나 확장이 가능하다. 토픽맵의 기본 기호(notation)는 SGML이다. 따라서 통합 대상이 되는 토픽맵 간에는 하나 이상의 토픽이 반드시 SGML형태로 이루어져야 한다. 이 때 통합이 가능하도록 완전하게 이루어진 토픽맵은 HyTime에서 정의하고 있는 'bounded object set(BOS)'를 이용하여 구별되어질 수 있다. 또한 토픽맵의 기본 기호는 WebSGML로 알려진 XML을 사용할 수 있다.

3.2 토픽맵의 구조

토픽맵 구조는 해당 자료의 기본틀과 위치정보, HyTime에 정의된 하이퍼링크 모듈로 이루어진다.

3.2.1 토픽맵 구조 형식

토픽맵 구조형식(Topic Map Architecture Form)에 정의된 요소는 topicmap과 added themes이다.

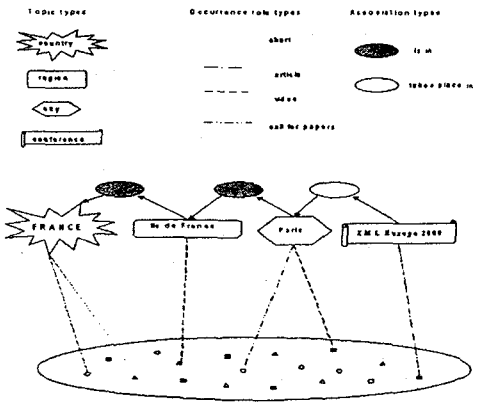
topicmap요소는 국제표준에 정의된 토픽맵 구조로 이루어진 모든 문서의 문서요소로 적용된다. 이에 비해 added themes 속성은 기부여된 토픽특성에 참조할 수 있도록 부가적인 주제를 부여하는 역할을 수행한다.

3.2.2 토픽 연결(topic link)

토픽 연결을 위한 요소는 토픽연결 구조요소(Topic Link Architectural Form)를 비롯하여 토픽명 구조요소(Topic Name Architectural Form), 토픽 대상물 구조요소(Topic Occurrence Architectural Form) 등 세가지 요소로 구성된다. 각각의 토픽구조 요소는 세부적인 요소를 갖고 있으며, 세부 요소들은 토픽맵 작성의 부수적이며, 설명적인 역할을 수행한다.

3.3 토픽맵의 모델 분석

토픽맵은 전통적인 색인화 도구인 분류표를 비롯하여 용어해설, 시소러스의 디스크립터간의 관계와 매우 유사한 구조를 갖는다. 예를 들면, 토픽 타입은 토픽맵 사이에서 유사 토픽간의 클러스터링 과정을 거친 후의 토픽군을 의미한다. 이 때 하나의 토픽 타입을 작은 토픽맵으로 정의할 수 있다. 이러한 토픽타입은 그 내부적으로 토픽간의 관계가 설정될 수 있으며, 토픽 타입사이에도 연관관계나 계층관계가 설정될 수 있다. 다음 <그림 1>는 토픽맵의 모델을 도식화한 예이다.



<그림 1> 토픽맵 모델 xmleurope2000

4. 시소러스의 지향점

시소러스는 정보검색시스템의 통제 주제분야에 따라 마이크로 시소러스(micro thesaurus)와 매크로시소러스(macro thesaurus)로 크게 구분할 수 있다. 마이크로 시소러스는 단일 주제를 개발 범위로 하였을 경우를 의미한다. 그러나 실제적으로 마이크로 시소러스를 한 주제에 대한 시소러스만을 의미하지는 않는다. 왜냐하면 대부분의 주제분야는 여러 주제가 합쳐져서 이루어져 하나의 주제분야를 이루는 것이 일반적이기 때문이다. 예를 들면, 경영학 시소러스라고 하여도 여기에는 순수 경영학 이외에 경제학이나 무역학, 외교 통상분야가 반드시 일정부분 포함되어 있기 때문이다. 한편 정보의 양이 급증함에 따라 시소러스의 형태도 마이크로 형태로 세분화되며 특정 영역의 정보수집을 위한 이용도구로 사용되는 것이 보편화되고 있다.(남영준 2002) 왜냐하면, 선진 각국에서 개발·활용되는 시소러스는 특

정영역의 전문자료의 검색만을 위한 시소러스가 대부분이기 때문이다. 각 시소러스에 수록된 디스크립터의 수가 제한적인 것도 대용량 시소러스의 무용성을 입증하는 것이다.

4.1 시소러스의 제한점

시소러스의 효율과 활용방안에 대한 연구는 오래전부터 이루어졌다. 이러한 연구는 전자자원이라는 방대한 규모의 데이터를 정보관리기관과 이용자들이 직면함에 따라 다음과 같은 제한점에 직면하게 되었다.

1) 정보확장을 수용할 수 있는 신규용어의 한계
2) 검색대상의 대용량화에 따른 재현율의 불필요한 향상
3) 적정한 정도율 확보의 어려움
4) 유사시소러스(외국어시소러스 포함)와의 통합의 제한점

이러한 제한점은 시소러스의 단점 측면보다 정보의 폭발시대에 급증하는 정보행태라는 외부적 요인에 크게 좌우되고 있다. 이러한 점을 극복하기 위해서는 시소러스의 유지관리를 유연하게 수행할 필요성이 있다. 예를 들면, 신규개념이나 신규용어의 수용을 자유롭게 유지하며, 불필요한 용어나 유용성이 극히 저하되는 디스크립터의 삭제나 축소를 용이하게 할 수 있도록 한다. 이러한 것은 매우 이론적인 것으로써 기존에 부여된 색인어로써 디스크립터에 대한 관리와 적정한 규모의 디스크립터를 유지해야하는 현실과는 괴리감이 발생할 수 있다. 즉 시소러스 구축에 따른 기본원칙은 디스크립터의 안정성과 적절한 규모유지로 대별할 수 있기 때문에, 대용량 데이터베이스 시대에 이 두가지 원칙을 수용하기 어려운 실정이다. 이러한 상반된 것을 극복하기 위해서는 매크로 시소러스로의 지향이 필요하다. 검색대상의 정보의 양이 급증하는 것은 해당분야의 연구자들이 급증하고 이는 궁극적으로 연구분야가 심화되고 확대되었음을 의미한다. 따라서 학문의 분화추세에 맞추어 시소러스도 매크로 형태로 개발되어야 한다. 또한 디스크립터의 자유로운 유지관리를 위해 개념어 위주의 디스크립터를 수용해야 한다.(남영준, 이두영 2004)

4.2 토픽맵 개념의 시소러스 개발

토픽은 인간이 용어로 표현될 수 있는 개념은 모두 가능하다. 일반적으로 토픽은 사람을 비롯하여, 사물, 개념, 의미 등 실제 존재하는 것, 또는 특정 속성이나 어떤 의미 등이 될 수 있다. 보통 명사형으로 표현된다. 즉 특정 문서의 토픽은 그 문서의 작성자가 나타내고자 하는 주제를 표현할 수 있는 단어들로 구성된다.(정호영)

토픽맵의 알고리즘을 사용할 경우, 시소러스와 토픽맵간의 관계는 다음과 같은 매칭으로 설명할 수 있다.

- ○ 토픽 / 디스크립터
- ○ 스코프 / 대등관계(다언어포함)
- ○ 연관관계와 토픽타입 / 연관관계 및 계층관계
- ○ 대상물 / U

특히 대상물은 특정 시소러스의 검색대상이 되는 검색대상물의 위치(즉, 적합정보)로 설명될 수 있다. 현실적으로 검색대상의 건수가 수작업으로 처리할 수 있는 범위를 넘어서면 모든 작업은 기계를 통해 이루어질 수 없다. 따라서 U는 디스크립터로 채택된 용어와 연관있는 문서의 집합이다. U의 규모가 커진다는 것은 고정된 정도율내에서 재현율이 높아지는 것을 의미하며, U의 규모가 낮다는 것은 재현율이 낮다는 것을 의미한다.

본 연구에서 제안하는 것은 앞의 4가지 요소를 시소러스에 적용(layout)하여 시소러스 디스크립터에 대한 개념의 안정성과 디스크립터의 가변성을 인정하는 유지의 용이성을 충족시키는 것이다. 또한 연관관계는 패싯개념을 적용할 수 있기 때문에 좀 더 세부적인 관계를 표현할 수 있다. 특히 연관관계는 다음과 같이 기호화함으로써 디스크립터간의 의미적 관계성을 표현할 수 있다.

파인애플 is a member_of 제주도특산물
특정인B bought a fruit, 망고
특정인A manage 제주도특산물

즉, 이때 제주도 특산물은 파인애플이나 망고보다 범용성이 있으며, 해당 용어군에서 개념적인(무형) 특성을 갖고 있다. 따라서 특정 상황에서 미래의 새로운 과일이 망고를 대치하여도 무형적인 특성인 제주도 특산물을 대치할 수 있는 개념은 없어 개념의 하이레벨의 변동은 없고 컨텐츠레벨의 파인애플과 망고와 같은 용어는 시의적절하게 수정될 수 있다.

한편 이러한 관계 표시는 기계를 통해 특정인 A와 파인애플간의 관계성을 연관짓고, 이를 표현하는 것이다. 즉, 특정인 A는 제주도 특산물을 관리하는 사람이고, 제주도 특산물에 파인애플이 포함되면 특정인 A와 파인애플은 관계성을 갖는 것이다.

5. 결 론

일반 시소러스의 디스크립터는 특정 영역의 자료를 검색할 수 있는 후조합 색인언어의 특성을 갖고 있으나, 토픽 맵은 색인된 특정 용어에 배정된 정보리소스도 검색하는 종합적인 검색도구이다. 따라서 토픽 맵은 모든 분야를 수용하는 것보다 특정 분야의 개념 구조화에 상대적으로 유리하다.

본 연구에서는 이 점에 착안하여 시소러스의 디스크립터 구조화 과정을 토픽맵의 구축 알고리즘을 이용하여 특정 영역의 마이크로 시소러스 구축 알고리즘을 제안하였다. 본 연구를 통해 제안된 것은 토픽맵의 기본요소를 시소러스의 관계 표시로 적용하는 것으로써 다음과 같이 요약할 수 있다.

○ 토픽 / 디스크립터
○ 스코프 / 대등관계(다언어포함)
○ 연관관계와 토픽타입 / 연관관계 및 계층관계
○ 대상물 / U

특히 대상물 U는 디스크립터와 연관된 정보원을 의미한다. 이는 웹상에서는 URL과 같은 소재정보를 비롯한 원문링크정보일 수 있으며, 전통적인 검색환경에서는 소장 및 위치정보의 역할을 수행한다.

이상과 같이 토픽맵 알고리즘을 시소러스 구축에 활용하여 전통적인 시소러스의 디스크립터 유지관리의 편의성과 개념의 안정성을 모두 확보할 수 있었다.

<참고문헌>

고유미. 2005. 「토픽맵 기반의 특허정보 서비스를 위한 시스템 구축에 관한 연구 : 항채이용기술 분야를 중심으로」, 숙명여자대학교 대학원 석사학위논문

고세영. 2003. 「토픽맵을 이용한 이 기종 상품분류체계 온톨로지 통합에 관한 연구」, 숙명여자대학교 대학원 석사학위논문

남영준. 2002. 「고속철도 시소러스 개발」. 쓰리소프트.

남영준, 이두영. 2004. 「로그데이터를 이용한 디스크립터의 외형적 특성 분석. 정보관리학회학술대회논문집」. 11:61-6

정호영 외. 2003. XTM 기반의 지식맵. 「데이터베이스연구」. 19(1) : 38-4

정원규. 2003. 벤담-도덕 및 입법의 원리 서설, 「철학사상」 제2권 제8호 별책

이은아. 2003. 「XML 토픽맵(XTM)을 이용한 시맨틱 네비게이션 시스템 구현」. 석사학위논문, 숙명여자대학교 대학원.

정준원. 2003. 「XML 기반의 지식맵 캐쉬 기법」. 석사학위논문, 서울大學校大學院.

유우종, 김진우, 권주홈. 2004. 「워드넷 온톨로지를 이용한 토픽맵 메핑」. 한국정보과학회 학술발표논문집:175-177.

ISO/IEC. 2002. ISO/IEC 13250:Topic Map

Le Grand, Soto, M.. 2000. "Information Management Topic Maps Visualization", XML Europe 2000, Paris, France.