특 집

정보검색에서의 시각화 기법

김 종 영, 오 희 국
한양대학교 전자계산학과

I. 소 개

최근 정보 저장 기술의 발전과 컴퓨터 네트워크의 확산으로 많은 양의 정보가 우리 주위에 존재하며 또한 이를 각자의 목적에 활용할 수 있는 기회가 계속적으로 증가하고 있다. 월드와이드웹이나 전자도서관은 이러한 흐름을 단적으로 표현해 주는 대표적인 예이다. 많은 정보로부터 사용자에게 쓸모있는 정보를 어떻게 효율적으로 끌어낼 수 있을까에 대한 고민이 정보검색의 동기가 된다. 정보검색은 데이터가 축적된 저장소에서 사용자의 정보요구에 적합한 대상을 선택하기 위한 일련의 과정이라고 말할 수 있다. 과거에는 정보검색의 대상이 주로 텍스트 문서였지만 최근에는 이미지, 소리, 동영상 등과 같은 멀티미디어 정보에 대한 관심이 증가하고 있다.

멀티미디어 기술의 발달로 검색 대상이 멀티미디어 정보로 옮겨가는 것과 함께, 이와는 조금 다른 이유에서 정보검색분야에 있어서의 시각화(visualization) 기법에 대한 관심이 최근 점점 높아지고 있다. 이는 크게 다음의 세 가지 이유에서 기인한다. 첫째는 정보검색 환경이 대용량의 데이터를 대상으로 한다는 것이다. 이로 인해 검색결과가 너무 많거나 부정확해서 사용자는 시스템이 제시하는 결과에 대해서 또 다른 검색을 하지 않으면 안되는 불가피한 상황을 맞이하고 있다. 둘째는 정보 검색 환경이 대상으로 하는 데이터의 전체적인 구조와 상호 관계를 사용자에게 효과적으로 제시할 필요가 있다는 점이다. 마지막으로 사용자는 자신의 정보요구(information need)를 정확히 질의로 표현하는데 있어서 어려움을 가진다는 것이다. 정보요구를 합당한 질의로 정확히 표현하기 위해서 사용자는 데이터 군(collection)의 특징과 정보검색이 제공하는 질의의 형태 등의 정보에 대해서 알고 있어야 한다. 시각화 기법은 이러한 문제점을 보완하고 보다 효율적인 검색 환경을 사용자에게 제공하기 위한 것으로 크게 질의 입력과 검색 결과 출력의 두 가지 단계에 적용한다.

정보검색에 있어서의 시각화 기법은 제한된 디스플레이 영역 내에서 다양한 형태의 정보 공간을 표시해야 하는 부담을 가지고 있다. 질의 입력 단계에서의 시각화는 사용자의 다소 관념적인 지식 공간을 적절히 나타낼 수 있어야 하며, 따라서 사용자가 질의를 효과적으로 표현할 수 있도록 해야 한다. 출력 단계에 있어서의 시각화 역시 복잡한 정보 공간을 사용자에게 효과적으로 제공해야 한다. 결과적으로 정보검색에 있어서의 시각화 기법은 제한된 디스플레이 공간을 어떻게 잘 활용하는가에 성공과 실패가 결정된다고 말할 수 있다.

II. 시각화의 분류

시각화 기법은 사용하는 데이터의 종류와 수행연산에 따라 분류할 수 있다[1, 2, 3]. 사용하는 데이터의 종류에 따른 분류는 다음과 같다.

- 1차원 데이터
- 2차원 데이터
- 3차원 데이터
- 트리 구조 데이터

● 네트워크 구조 데이터

1차원 데이터는 텍스트 데이터를 주로 말하며 텍스트 문서, 프로그램 소스 코드 등과 같은 선형 데이터가 여기에 해당된다. 1차원 데이터의 시각화에 있어서의 고려 사항은 색깔, 크기, 폰트 등이 있다. 2차원 데이터는 지형 정보를 나타내는 지도 또는 건축물의 설계도와 같은 데이터 타입이 여기에 해당된다. 2차원 데이터는 주로 직사각형과 같은 다각형을 이용하여 나타내며 2차원 데이터에 대한 사용자의 관심은 인접 개체의 검색, 개체들 사이의 경로 검색, 특정한 조건을 만족하는 개체의 개수 파악 등이 있다.

3차원 데이터는 건축물이나 분자 구조와 같은 실세계의 개체를 말한다. 사용자의 관심은 3차원 데이터의 상하와 내외부로의 탐색이다. 3차원 데이터를 이용한 시각화 기법에서 고려해야 할 사항은 사용자가 자신의 위치를 항상 파악할 수 있도록 해야 한다는 것이다. 트리 구조는 개체들 사이의 계층 구조를 표현하는데 유용하게 사용할 수 있다. 트리 구조에서는 루트를 제외한 모든 노드들은 바로 상위 노드로의 링크를 가지고 있다. 네트워크 구조 데이터는 개체들 사이의 상호관계를 나타내는데 있어서 트리 구조가 가지는 단조로운 구조의 단점을 보완할 수 있다. 네트워크 구조는 하나의 개체가 임의의 개수의 링크를 가지게 하는 방법을 통하여 개체들 사이의 상호관계를 표시한다.

데이터에 대한 수행 연산에 따라 시각화 기법을 분류하면 다음과 같다.

- Overview
- Zoom
- Filter
- Details-on-demand
- Relate
- History
- Extract

수행 연산에 있어서 Overview는 전체 정보 공간의 개요를 얻기 위한 연산이며, Zoom은 Overview와 반대되는 연산이다. Zoom은 관심이 되는 개체를 확대시키며 대부분의 경우 Zoom 배율과 Zoom 초점 기능을 제공한다. Filter는 관심의 대상에서 제외되는 개체를 제거하는 연산이며, Details-on-demand는 최종적으로 관심의 대상을 선택한 경우 그 대상의 세부적인 사항에 대한 정보를 얻고자 할 때 사용되는 연산이다. Relate는 개체들 사이의 상호 관련성을 알아보고자 하는 연산이다. History는 사용자의 연산을 기록하고 저장함으로써 나중에 다시 조회할 수 있도록 사용자의 편의를 제공한다. Extract는 사용자가 원하는 결과를 최종적으로 얻었을 경우 그 결과를 출력하는 방법으로 인쇄와 저장 등이 있다.

앞에서 나열한 데이터나 연산의 종류 외에도 정보검색에 있어서의 시각화 기법은 어떤 속성값을 주요 표현의 대상으로 하느냐에 따라 달라지며, 검색 시스템이 채택하고 있는 모델에 영향을 받는다. 문서 검색의 경우, 표현의 대상이 되는 속성으로는 순위, 적합성 수치 정보, 문서 내의 키워드 분포 정보, 문서의 길이, co-occurrence 정보 등이 있다. 검색 시스템이 채택하고 있는 모델은 곧 검색 대상의 내부적 표현을 뜻하며 대표적으로 불리언 모델, 클러스터 모델, 벡터 공간 모델 등이 있다. 특히 불리언 모델을 채택한 경우의 시각화 기법은 결과의 출력보다는 사용자가 질의를 효과적으로 만들 수 있도록 도와주는 역할을 한다.

다음 장에서는 1차원 데이터 타입인 TileBars와 2차원 데이터 타입인 InfoCrystal, 3차원 데이터 타입인 Cat-a-Cone, PRISE 사용자 인터페이스에 대해서 소개한다. 이들 시스템은 정보검색 분야에서 이슈가 되는 문서의 길이, 키워드 분포 정보, 계층 구조, 클러스터링, 불리언 질의 형성, 문서 공간의 정보들을 이용하여 효과적으로 사용자에게 제시함으로써 시각화의 목적을 달성하고 있다.

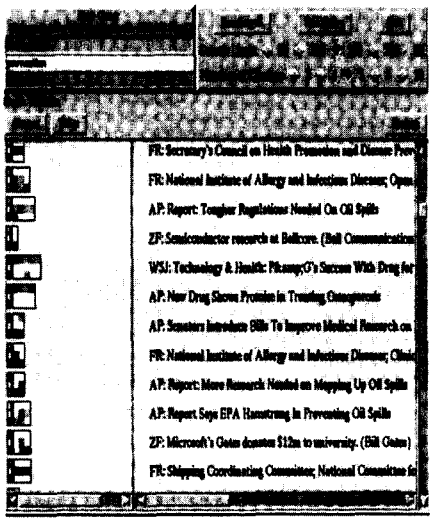
III. 정보검색 분야의 시각화 사례

1. TileBars

과거의 정보검색 환경은 주로 제목이나 개요 수준의 문서를 그 대상으로 하였지만 오늘날의 전문

(full-text) 검색 환경에서는 문서의 구조와 문서 간의 상호 관계를 중요시한다. TileBars는 문서의 구조 정보 중에서 문서 내의 키워드 분포 정보를 시각화한다. 키워드의 분포 정보가 중요한 까닭은 문서의 길이를 고려할 경우 분명해 진다. 짧은 길이의 문서와 긴 문서 양쪽에 포함되는 키워드는 각각의 문서의 순위에 끼치는 정도가 다르다는 사실은 과거의 제목이나 개요 수준의 문서를 검색 대상으로 했던 시절에는 고려되지 않았다.

TileBars는 사용자에게 검색된 문서의 상대적인 길이, 질의에 포함된 키워드의 상대적인 빈도수와 문서 사이의 분포 정보 등을 시각적으로 제공한다[4]. 그림 1[1]에서 각각의 사각형은 문서를 나타내며, 사각형 내의 타일 모양의 작은 정사각형은 문서를 다수의 단락으로 나누었을 때 한 단락을 의미하는 TextTile을 나타낸다. 즉, 열은 하나의 단락을 의미하며 행은 특정한 키워드를 의미한다. 행과 열의 교차점에 있는 TextTile의 색깔이 검을수록 그 단락에 특정 키워드가 많이 포함되어 있다는 것을 의미한다. 따라서 같은 열에 있는 TextTile내에 키워드들이 같이 포함되어 있으면 이 키워드들은 동일한 소 주제를 나타낸다고 말할 수 있다.

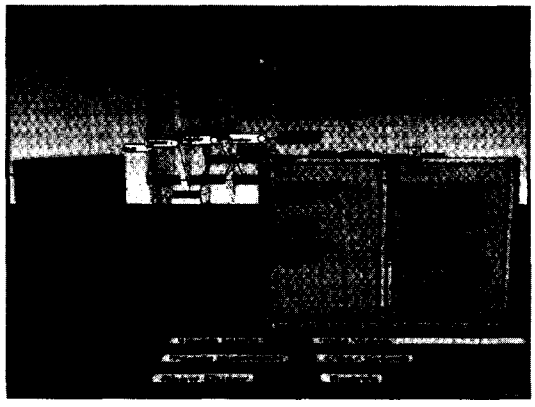


〈그림 1〉 TileBars[4]

2. Cat-a-Cone

오늘날 대부분의 문서군들은 특정한 주제에 의한 분류되어 있다. 각각의 범주는 분류 표식(category label)을 이용하여 나타낸다. 월드와이드웹에 있어서 대표적인 검색 사이트인 Yahoo는 이러한 주제별 분류 표식을 제공한다. 이러한 분류 표식은 문서군에 따라 다르지만 대체로 그 용도는 사용자가 질의를 형성하는 작업을 도와준다. 하지만 대부분의 경우에 있어서 이러한 분류 표식 사이에 존재하는 계층구조를 효과적으로 탐색하고 선택할 수 있게 하는 인터페이스를 제공하고 있지 못한 것이 사실이다.

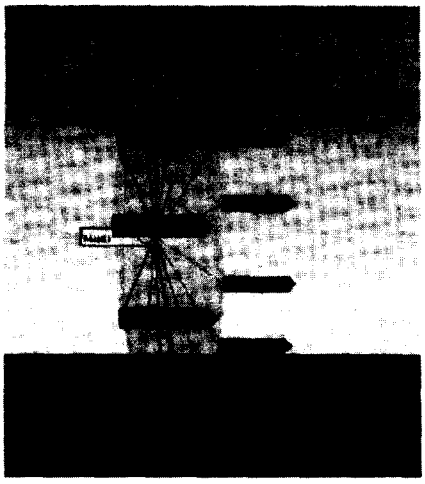
XEROX에서 개발된 Cat-a-Cone은 문서군의 탐색과 검색을 지원하기 위하여 개발된 인터페이스이며 3차원 애니메이션 기법을 도입하여 제작되었다[5]. Cat-a-Cone은 하나의 문서에 여러 개의 분류 표식이 할당되어 있는 경우에 문서를 분류 계층구조로부터 분리한다. 이렇게 함으로써 보다 나은 검색과 디스플레이를 제공할 수 있다.



〈그림 2〉 Cat-a-Cone 인터페이스[5]

그림 3은 의학 관련 문서군인 MEDLINE 내에서 가슴 부위에 발병하는 암에 대한 부분 문서군을 검색하기 위한 초기 계층 구조를 나타내고 있다.

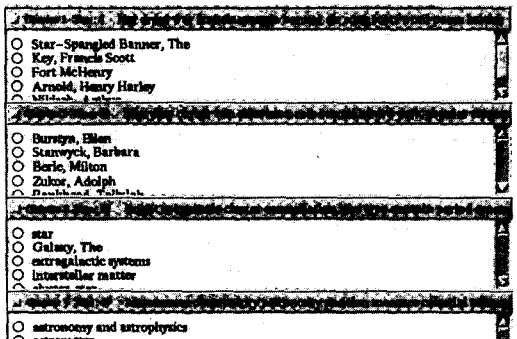
1) 본 논문에 삽입된 그림은 원 저자의 허가를 받아 게재함.



〈그림 3〉 Cat-a-Cone 초기 검색화면[5]

3. Scatter/Gather

정보 검색에서 문서 클러스터링은 다음의 두 가지 이유로 인해 그다지 널리 사용되지 못하고 있다. 첫째는 문서의 개수가 많아지면 많아질수록 문서의 클러스터링에 소요되는 시간이 급격하게 증가한다는 것과, 둘째로 클러스터링 방법이 검색의 질을 그다지 크게 향상시키지 못한다는 것이다. 하지만 기존의 클러스터링 방법은 전체 시스템에서 차지하는 비율로 볼 때 주된 방법이라기 보다는 검색 시스템을 부분적으로 도와주는 의미에서 사용되었으며, 클러스터링 알고리즘 자체도 뛰어 나지 못한 게 사실이었다[6].



〈그림 4〉 Scatter/Gather[7]

Scatter/Gather는 대용량의 문서군을 탐색하기 위해 클러스터링 방법론을 채택하고 있다. 또한 빠른 문서 클러스터링 방법을 도입하여 시간 복잡도(time complexity)를 기존의 2차(quadratic)에서 선형(linear)으로 향상시켰다[7]. Scatter/Gather는 초기 검색 단계에서 문서를 클러스터라고 불리우는 몇 개의 문서 그룹으로 나눈다. 이때 시스템은 각각의 클러스터에 대한 요약 정보를 제시하며 사용자는 이를 통해 검색 후보를 선택하게 된다. 선택된 클러스터들은 다시 한번 모아져서 하위 문서군을 형성하며, 몇 개의 클러스터로 다시 나누어서 사용자에게 검색 후보를 선택하게 한다. 사용자는 이러한 과정을 반복적으로 진행하여 원하는 결과를 얻을 때까지 계속한다.

4. InfoCrystal

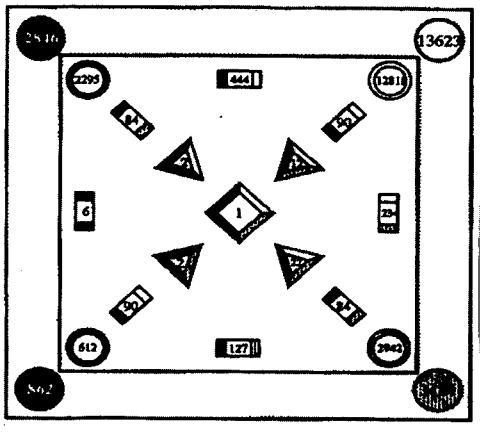
정보 검색에 있어서 불리언 질의의 단점은 때때로 너무 많거나 또는 너무 적은 숫자의 결과를 사용자에게 제시한다는 것이다. 논리 연산자 AND를 사용하여 형성된 질의는 너무 적은 숫자의 결과를 주며, 반대로 논리 연산자 OR를 사용하여 형성된 질의는 너무 많은 숫자의 결과를 준다. 이런 경우 사용자는 얻은 결과를 확장하거나 축소하기를 원한다.

MIT에서 개발된 InfoCrystal은 불리언 질의를 효과적으로 처리할 수 있도록 질의를 시각적으로 구성하도록 해주며 동시에 검색 결과를 시각적으로 나타내기 위해서 개발되었다[8]. 전반적인 InfoCrystal의 기능은 다음과 같이 요약해 볼 수 있다. 첫째, 전체적인 개요를 유지하는 범위 내에서 정보 공간을 다양한 차원에서 탐색할 수 있다. 둘째, 사용자는 질의의 원래 형태를 유지하면서 다양한 방법으로 질의를 변화시키고, 그에 따른 결과를 관찰하는 것이 가능하다. 셋째, 사용자는 검색 과정에 있어서 시스템으로부터 필요한 시각적인 피드백을 받는다. 이렇게 함으로써 사용자는 나머지 검색 과정을 안내받게 된다. 넷째, 사용자는 질의를 시각적으로 구성할 수 있으며, 검색 방법을

불리언 또는 벡터 검색 방법과 같이 다양하게 선택할 수 있다.

InfoCrystal은 2차원 공간의 디스플레이에 여러 개의 검색 키워드 사이에 존재하는 관계를 시각화하기 위하여 벤 다이어그램을 이용한다. N개의 검색 키워드에 대해서 InfoCrystal은 2^N-1개의 논리 조합(boolean combinations)을 만들어내며, 각각의 경우는 Interior Icon이라고 불리우는 아이콘에 의해서 표시된다. 이 아이콘을 표시하는 방법에는 정성적인(quantitative)방법과 정량적인(qualitative)방법 두 가지가 있다. 정성적인 방법은 각각의 Interior Icon들이 입력과 어떻게 연관되어 있는가를 나타내는데 사용된다. 정량적인 방법은 각각의 Interior Icon들이 몇 개의 문서와 연관되어 있는가를 나타내는데 사용된다. 이 때 Interior Icon의 색깔과 모양을 통해 문서의 적합성을 표시한다.

InfoCrystal 환경에서 사용자는 불리언 연산자나 괄호 등을 사용할 필요가 없으며 단지 검색 키워드들 사이의 관계만 파악하고 있으면 된다. 이 때, 사용자와 시스템은 직접적인 상호작용을 통해서 질의를 시각적으로 형성하게 된다. 그림 5는 네 개의 검색 키워드를 사용했을 경우의 질의를 시각적으로 나타내고 있다.



〈그림 5〉 InfoCrystall[8]

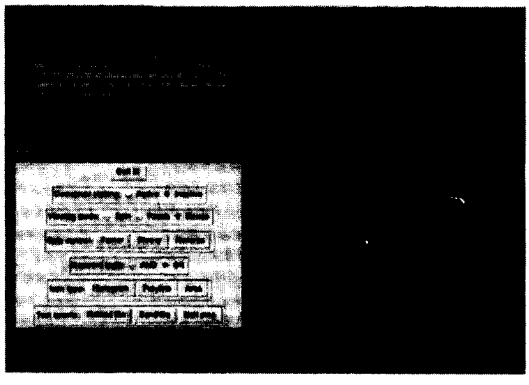
5. PRISE 사용자 인터페이스

대부분의 검색 시스템은 문서의 제목을 선형으로 배열하여 결과를 제시한다. 하지만 그러한 방법은 여러 가지 문제점을 가지고 있다. 첫째로 얻은 결과가 상당히 많은 경우 사용자는 그 결과를 일일이 살펴보아야 한다는 점이다. 둘째는 얻은 문서들이 서로 어떠한 연관성을 가지고 있는지 알기가 어렵다는 점이다. 예를 들어, 사용자가 반환된 리스트에서 질의에 적합한 문서를 찾았을 경우에 관련된 다른 문서를 찾기가 어려울 수 있다. 마지막으로 검색 시스템이 제시한 검색 순위가 사용자가 기대하는 순위와 일치하지 않을 수 있다는 점이다.

PRISE 검색 시스템은 앞에서 언급한 문제점을 갖고 있는 시스템으로 NIST(National Institute of Standards and Technology)에서는 이러한 문제점을 해결하기 위해 향상된 인터페이스를 개발하고 있다[9]. 다음은 PRISE 검색 시스템에 이식된 인터페이스와 시각화 방법에 대해서 기술한다.

(1) Document Spiral

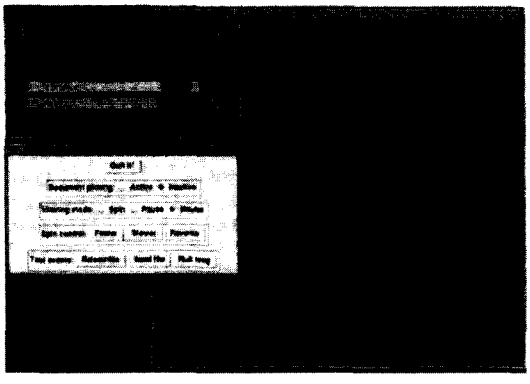
이 방법에서는 문서를 나선형의 곡선 상에 배열시킨다. 사용자의 질의에 적합한 문서일수록 나선형 곡선의 중심에 위치하게 되며, 반대의 경우 곡선의 중심으로부터 멀어지게 된다. 문서는 아이콘으로 표시되며, 같은 적합성을 가진 문서들이 겹치는 것을 피하기 위해 따로 따로 표시한다. 또한 문서 밀도라는 값을 조절함으로써 곡선 상에서 문서간의 거리를 조절할 수 있다. 이 방법의 장점은 검색 결과 내에서의 문서 분포를 쉽게 알 수 있다는 점이다. 즉 곡선 상의 문서 아이콘이 균일하게 분포하고 있는지 또는 중심에 몰려 있는지 등의 사실을 시각적으로 알 수 있도록 한다.



〈그림 6〉 Document Spiral[9]

(2) Three Keyword Axes Display

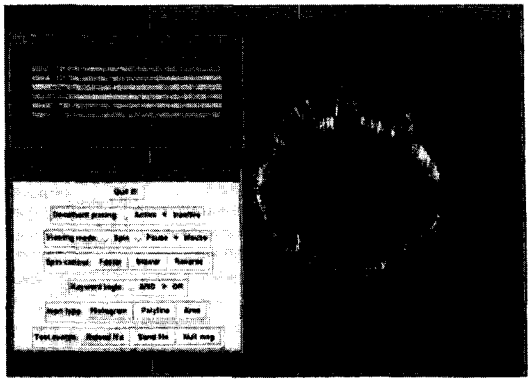
이 방법은 순위 정보를 직접적으로 표시하는 대신 문서를 3차원 공간에 나타낸다. X, Y, Z축은 각각 하나 이상의 키워드를 할당받는다. 이 때 3차원 공간상에서의 문서의 위치는 질의 내에 포함된 키워드가 전체적인 문서의 순위에 얼마만큼의 영향력을 미치는가에 따라 동적으로 결정된다. 사용자가 각각의 축에 대해서 다른 키워드들을 할당한 경우, 문서의 위치는 자동으로 변환된다. 문서는 좌표 공간에서 역시 아이콘으로 표시가 되며 문서의 적합성 정도는 아이콘의 색깔에 의해서 구분이 된다. 이 방법의 장점은 사용자가 시스템과 직접 대화함으로써 문서 내부에서의 질의 키워드들에 대한 분포 정보를 쉽게 파악할 수 있다는 점이다. 하지만 이 방법의 단점은 세 개 이상의 키워드가 사용되는 경우 각각의 X, Y, Z축 가운데 몇 개는 두 개 이상의 키워드가 할당되므로 각 키워드가 문서의 위치에 어떠한 영향을 미치는지 알기 어렵다는 점이다.



〈그림 7〉 Three-Keyword Axes Display[9]

(3) Nearest Neighbor Sequence

Nearest Neighbor Sequence 방법은 의미적으로 유사한 문서들을 모아 클러스터를 형성하는 방법이다. 이 때 문서는 키워드 영향력 벡터로 나타낸다. 임의의 두 문서가 의미적으로 얼마나 가까운가를 나타내는 의미적 거리(semantic distance)는 각 문서의 키워드 영향력 벡터를 이용하여 계산하며, 유클리디언 거리, 또는 두 벡터 사이의 각도를 고려함으로써 구한다. 이렇게 계산한 거리와 각도



〈그림 8〉 Nearest Neighbor Sequence[9]

에 따라 각 문서들은 의미 공간(semantic space)에 배치된다.

IV. 결 론

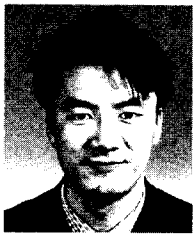
본 논문은 정보검색 분야에서 사용되는 대표적인 시각화 기법에 대해서 기술하였다. 시각화 기법의 공통적인 이점은 검색 결과의 내용을 유지하면서 검색 결과의 개요를 사용자에게 다양한 형태로 제공한다는 점이다. 또한 검색 결과에 대해 사용자가 검색 시스템과 상호 작용할 수 있는 방법을 제공한다는 점이다. 하지만 제한된 2차원 공간에 많은 수의 개체를 표시해야만 하는 현실을 고려할 때, 보다 효율적이며 적절한 시각화 방법이 요구된다. 특히 사용자와 검색 시스템 사이의 relevance feedback 과정 역시 시각화의 대상이며, 앞으로 연구가 계속되어야 할 분야이다. 이 밖에도 질의 형성에 관한 시각화 기법도 앞으로의 연구과제이다.

참 고 문 헌

[1] Shneiderman B., The Eyes Have It: A Task by Data Type Taxonomy of Infor-

mation Visualizations, Proceedings of IEEE Symposium on Visual Languages '96, IEEE, 336-343, Los Alamos, CA, 1996.
[2] Cleveland, W., Visualizing Data, Hobart Press, Summit, NJ, 1993.
[3] Korfhage, R., To See or not To See -- Is that the Query?, Communications of the ACM 34:134-141, 1991.
[4] Hearst, M., TileBars: Visualization of Term Distribution Information in Full Text Information Access, Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems (CHI), Denver, CO, 1995.
[5] Karadi, C. and Hearst, M., Cat-a-Cone: An Interactive Interface for Specifying Searches and Viewing Retrieval Results using a Large Category Hierarchy, Proceedings of the 20th Annual International ACM/SIGIR Conference, Philadelphia, PA, 1997.
[6] Croft B., A Model Of Cluster Searching Based On Classification. Information Systems, 5:189-195, 1980.
[7] Karger, D., Hearst, M., and Pedersen, J., Scatter/Gather as a Tool for the Analysis of Retrieval Results, Working Notes of the AAAI Fall Symposium on AI Applications in Knowledge Navigation, Cambridge, MA, 1995.
[8] Spoerri, A., InfoCrystal: a Visual Tool for Information Retrieval & Management, Proceedings of the Second International Conference on Information and Knowledge Management, Washington D.C., 1993.
[9] http://zing.ncsl.nist.gov/~cugini/uicd/viz.html.

저 자 소 개



金 鍾 英

1971년 9월 7일생, 1996년 2월 한양대학교 전자계산학과 학사, 1998년 2월 한양대학교 전자계산학과 석사, 1998년 3월~현재 한양대학교 전자계산학과 박사과정, 〈주관심 분야 : 정보검색, 인공지능〉

吳 熙 國

1960년 1월 12일생, 1982년 2월 한양대학교 전자공학과 학사, 1989년 8월 Iowa State University 전자계산학 석사, 1992년 11월 Iowa State University 전자계산학 박사, 1993년 3월~1994년 2월 전자통신연구소 선임연구원, 1994년 3월~현재 한양대학교 전자계산학과 조교수, 〈주관심 분야 : 정보검색, 분산시스템〉