4. 데이터,정보,지식 그리고 구조적 의미

2010 데이터로 표현하는 세상 요약본
고려대학교 김현철 교수
hkim64@gmail.com

구조적 정보와 의미적 정보, 그들간의 관계

"정보"는 우리에게 무엇을 주는가? "정보"는 우리에게 "의미"을 준다. 우리가 "데이터"라고 이야기 할 때에는 "의미"적인 면을 별로 생각하지 않는다. 그 데이터는 정보를 만들기 위한 기초 소스로서 역할을 한다. 그리고 의미라는 것은 주관적인 것이어서 사람마다, 그리고 시대마다 달라지기 때문에 여기서 우리는 데이터와 정보라는 용어를 굳이 칼로 자르듯이 구분하는 것이 무의미할 것이라는 생각을 하게 된다. "김현철"이라는 단어도 어떤 사람에게는 "데이터"에 불과하고, 어떤 사람에게는 "정보"의 의미를 갖기 때문이다. 본문을 읽는데 한 가지 주의할 점은 본문에서는 데이터와 정보라는 용어를 혼돈하여 사용하게 될 수도 있다는 것이다. 그 이유는 데이터와 정보는 상대적으로 서술이 되기 때문이다. 어떠한 의미 있는 "정보"도 그것을 이용하여 더 상위 레벨의 의미적 정보를 끄집어 낸다면, 그 상위 레벨의 정보의 관점에서는 하위 레벨의 정보를 데이터라고 생각할 수도 있기 때문이다. 마찬가지로, "지식(knowledge)"란 용어도 매우 상대적이다. 일반적으로 지식은 어떤 구체적인 특정한 정보라기 보다는, 아주 일반화 시킨 정보의 형태라고 생각할 수 있다.

'정보'의 전달은 즉 '의미'의 전달을 이야기 하는 것인데, 그러면 그 '의미'는 어떻게 표현이 되는 것일까. 다음의 그림을 보자. 두 개의 도시락 밥 위에 올려진 콩이 우리에게 주는 의미는 같은가. 똑같은 밥과 콩인데. 그 둘은 우리에게 다른 의미로 전달된다. 무엇이 그 의미를 만들었는가.

<u>시각적 구조</u>

위의 도시락의 예에서 왼쪽의 콩 배열이 그냥 '데이터'라고 한다면, 오른쪽 콩 배열은 '정보', 즉 '의미'를 담으려 한다. '하트'는 사랑이라는 것을 의미하는 심볼(기호)라는 것을 우리는 알고 있기 때문에, 콩의 배열을 그 기호를 시각적으로 보여 줌으로서 '사랑'이라는 의미를 전달하려 한 것으로 생각할 수 있다.

여기서 콩 하나는 정보의 기본 단위라고 말한다면, 정보의 기본 단위를 어떠한 형태로 '구조화'하여 의미를, 즉 '정보'를 표현하였다. 여기서의 그 구조는 시각적 구조이다. 만약에 그 정보단위들이 특정한 구조를 갖지 못했다면 '의미' 없는 데이터의 나열이 된다. 즉, 정보의 '구조'는 '의미'를 표현할 수 있다.

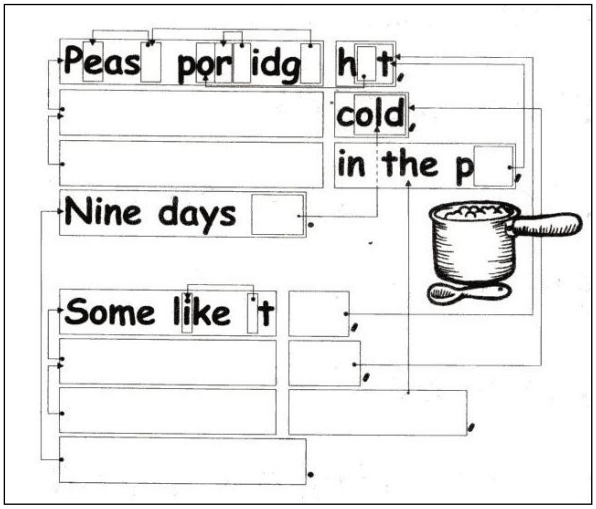
<u>순차적 구조: 단어, 문장, 문단</u>

정보단위를 "구조"화 함으로서 "의미"를 만들어 낼 수 있다. 간단한 문제이긴 하지만, 정보의 단위와 구조적 정보와 의미적 정보에 대한 이해를 하고 있는 것이며, 원하는 의미를 표현하기 위하여 정보단위들을 어떻게 구조변화를 하는지에 대한 알고리즘, 순차적 알고리즘까지 표현하고 있다. 여기서는 "의미"는 "순서적 구조"의 의하여 보여지고 있다. (물론, 구조에는 시각적 구조, 순서적 구조만 있는 것은 아니다).

이와 같이 우리 인간은 “본능적으로” 이렇게 정보를 다루는 (표현하고 구성하고 재 조합하는) 능력을 가지고 태어났다.
이렇게, 문장의 기본 구성 단위를 단어라고 정의를 하고, 그 기본 정보 단위의 구조를 조정함으로써 우리가 할 수 있는 일은 단지, 의미의 재 구성뿐만 아니라 우리가 이전 시간에 배웠던 정보의 압축(compression)에도 사용할 수 있다. 텍스트 압축에서 보았던 키워드 압축 방법이나 run-length 압축 방법이 그런 예에 해당된다. Huffman코딩은 단지 문자나 단어를 정보의 단위로 정의하여 그것을 대치하는 정보에서 더 나아가, 그 문자나 단어가 전체 문서에서 사용되는 빈도수를 그 문자나 단어의 정보요소에 포함을 시킴으로써 향상된 효율성을 보였다.

다음과 같은 압축의 예를 보도록 하자.

아래의 시는 단어와 문자들이 빠져 있다. 빠진 문자와 단어는 화살표가 가리키는 위치의 문자나 단어를 넣으면 완성된 시를 만들 수 있다.



〈 출처 : Computer Science Unplugged, Tim Bell, Etc. 1998 〉

언어를 정보단위로 ‘분해’하여 ‘재구조화’하는 과정은 새로운 ‘의미’를 부여하는 과정이다. 무작위로 재구조화 한다면 그것은 의미 없는 ‘소리’에 불과할 것이며, 같은 정보 단위를 가지고도 다른 구조로 조합 함으로서 다른 의미를 만들어 내기도 한다.

정보를 구성하는 요소는 ‘내용’뿐만이 아니다. 내용은 기본적으로 갖추어야 하지만 그것을 더 좋은 정보로 만드는 것은 그것의 ‘구조화’이다. 구조화의 능력이 개인의 능력을 좌우한다. 책 ‘논리의기술’에서는 정보를 받아들이는 사람의 정보청취와 생각의 구조를 참조하여 그것에 맞는 정보의 구조를 만들어 그에게 전달할 것을 이야기 하고 있다.

문제 해결 지식의 구조와 의미

앞 절에서는 “정보”를 단위정보들의 구조화와 그리고 그 구조가 주는 의미적 정보에 대하여 이야기를 했다. 그리고 그 단위정보들은 콩, 문자, 단어와 같은 단편적인 정보(혹은 데이터)들이었으며, 그들의 구조화를 통하여 얻어진 의미적 정보도 대부분 사실적(fact) 지식에 가까운 예들이었다. 하지만 지식에는, 정보에는, 사실적 지식만 있는 것이 아니라, 절차적 지식, 개념적 지식 등도 있다. “문제 해결 지식”이라는 용어는 사실적/절차적/개념적 지식을 모두 통합한 형태의 지식을 사용하여 문제 해결에 사용한다는 것을 의미한다. 다시 말하면, 복잡한 문제 해결에 사용되는 지식은 그 하위 지식, 혹은 정보들의 구조에 의하는 것이고, 그 구조를 이루는 단위 정보 혹은 단위 지식은 단편적 단위가 아닐 수도 있다.

마치 축구에서의 포메이션(formation)과 비슷하다. 포메이션은 크게 '수비수의 수'에 따라 나뉜다. 이를테면, 스리백, 포백, 파이브백이다. 또한, 각각의 수비수의 수에, 미드필더와 공격수를 어떻게 배치하느냐에 따라 위와 같이 나뉜다. 이러한 포메이션은 상대팀에 따라서, 혹은 경기 상황에 따라서, 게임의 목적에 따라서 다르게 구성할 수 있다.

여기서, 정보의 단위는 각 개별 선수들이며, 포메이션은 그 단위 정보의 구조를 나타낸다. 그리고 그 구조는 이 축구팀의 의도하는 전술 (즉, 의미, 목적, 역할)을 보여준다. 구조에 따라서, 그 역할과 목적과 능력이 달라지게 되는 것이다. 정보의 구조, 혹은 지식의 구조는 문제 해결 능력을 좌우한다.

절차적 지식의 구조적 표현

라면을 "잘" 끓일 줄 아는가? "라면을 잘 끓인다"는 것은 문제해결의 지식이 필요하다. 그 중에 특히 절차적 지식이 필요한 것이다. 여기에서 정보의 단위는 무엇일까? 그 정보의 단위들을 어떠한 형태로 구조화 하여야 그 문제해결의 지식을 나타낼 수 있을까?
여기서 단위 정보(혹은 지식)은 각기 잘게 쪼개어진 task(or activity)로 나타낼 수 있다. 즉, "라면을 끓인다"라는 포괄적인 문제를, 잘게 분해하여 (decompose)하여 단위 task들로 만들고, 그 단위 task들을 구조화하면, 라면 끓이는 과정을 보여줄 수 있다.

〉라면 봉지를 뜯는다.
〉라면을 꺼낸다.
〉냄비에 물을 넣는다 (얼마만큼??)
〉물이 끓으면 라면을 넣는다.
〉김치를 넣는다.
〉계란을 넣는다.
등등…

이것은 각 단위 task들을 "순차적(procedural)" 구조로 표현한 것이다. 이 순서가 맞는가? 더 나뉘어질 수 없는 작은 단위의 task로 분해되어 있는가? 김치를 넣는 순서를 바꾸면 어떻게 되는가? 다른 맛의 김치가 될 것이다! 치즈를 넣으면?
이것도 정보단위의 구조화에 따라서 다른 라면이 만들어 지는 것을 보여주고 있다. 즉, "정보의 단위"를 정하고 그 정보의 단위들을 "어떻게 구조화" 하여 "내가 원하는 의미"를 만들 수 있도록 "일련의 작업, 즉 알고리즘"을 정의하는 일이 우리가 관심 있어 하는 일이다. 여기서의 구조는 절차적 구조를 말한다.

정보의 단위들

우리는 앞에서 다양한 여러 가지 정보를 어떻게 discrete (혹은 digital) data로 변환하느냐 (즉, computing이 가능한 형태의 data로 변환하느냐)에 대하여 학습하였다. 하지만, 분명한 것은, 우리가 배운 그 방법이 "정답"은 아니라는 것이다. 새로운 방법의 더 낳은 방법이 존재할 것이라는 것은 분명하다. 다만, 지금 그것을 모르고 있을 뿐이다. 우리가 학습했던 것은 그 동안 많은 사람들이 고민하고 연구하였던 것을 한번 review해보는 것이지 그리고 그것을 알기 때문에 우리가 더 좋은 새로운 방법을 고안하게 하고자 하는 것일 뿐이다.

우리가 알고 있는 정보나 지식은 구조화를 통하여 의미화 시키고 인코딩을 통하여 Computing Model로 만든다고 생각할 수 있다. 그리고 구조화와 인코딩을 위하여서는 우리의 정보나 지식이 단위정보, 혹은 단위지식의 형태로 분해되어야 한다고 하였다. 얼마나 다양한, 그리고 효율적인 단위정보를 선택하여 사용할 수 있는가는 그래서 매우 중요한 문제이다. 단위정보의 양과 질에 따라서 구조화의 질이 달라지게 되고 그것은 computing model의 질을 좌우하기 때문이다.
여기서 한가지 고려해야 하는 점이 있다. 그것은 바로 "매체(media)"의 역할이다. 우리 머리 속에 들어있는 "지식, 정보"는 매우 implicit한 형태로 들어 있으며, 그것은 "매체"를 통하여 (혹은 매체의 도움을 받아서) 변환되어 우리 뇌의 바깥쪽으로 나온다.
내가 전달하려고 하는 정보/지식의 종류에 따라 우리는 그것을 가장 잘 표현 해 낼 수 있는 매체를 사용한다. 그 매체는 그 정보/지식을 잘 표현, 전달 할 수가 있다는 가정을 하고 있다. 잘못된 매체의 선택은 우리가 의도하는 정보/지식을 전달하지 못할 수도 있다. 즉, 매체가 정보의 질과 양을 조절할 수 있다는 것이다. 매체는 때로는 우리가 의도하지 않는, 의도 하지 않았던 정보까지 꺼내 보일 때도 있다.
매체는 내 머리 속의 정보/지식을 끄집어 내여 표현하게 하는 프레임 역할을 하게 되는데, 이때 그 정보의 단위정보들을 "구조화"하는 특별한 역할을 한다. 각 매체마다 그 나름대로의 "기본 구조"를 가지고 있다. 우리가 어떤 정보/지식을 어떤 특정 매체에 얹으면 그 특정 매체의 기본 구조 틀에 우리의 정보/지식에 맞추어져서 구조화 될 수 있다.
언어는 문자(소리)들의 1차원 순차적 구조 틀을 가지고 있다. 또한 목소리를 가지고 있으므로 감정적, 정서적 정보를 담을 수 있는 틀을 가지고 있다. 노래는 언어적인 정보와 음의 높낮이, 리듬, 음질 등의 구조로 정보를 전달한다. 그림은 색깔과 질감과 이미지의 구조를 가지고 정보를 전달한다. 우리는 내가 전달하고자 하는 정보/지식을 가장 잘 표현(구조화)할 수 있는 매체를 사용하고자 한다.
"매체"속에는 엄청나게 많은 종류의 "정보"가 들어 있다. "매체(정보)"는 우리가 그것을 가지고 무엇을 하려고 하느냐에 따라서 다른 representation을 할 수가 있다. 먼저, "매체(정보)"의 관계를 예를 가지고 보도록 하자.

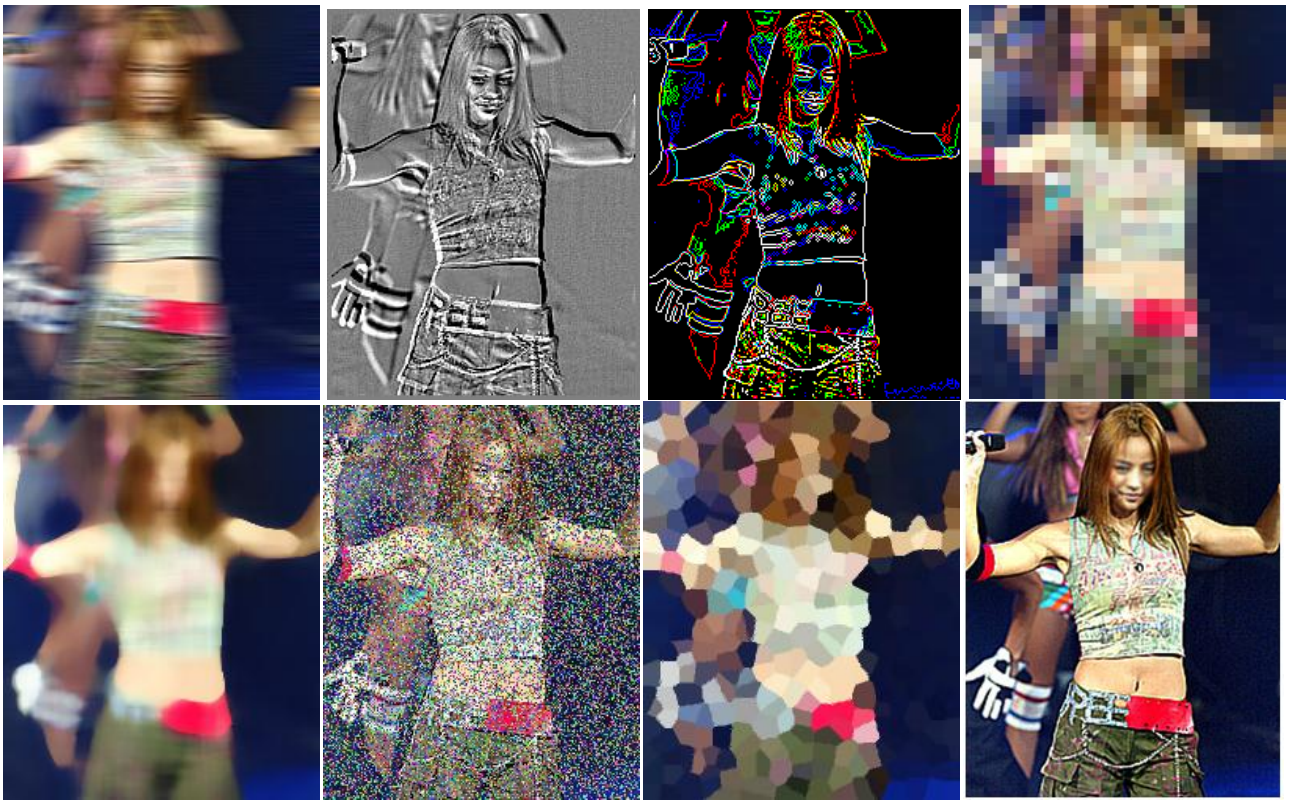
"언어"를 생각해보자. 말로 전달하려고 하는 정보 자체는 있는 것이고, 그것은 매체, 즉 목소리라는 매체에 붙어서 우리에게 전달된다. 그 말을 문자(즉 심볼)로 변환하면 우리는 매체를 벗겨내고 정보자체만 끄집어 낼 수 있다. 끄집어 낸 symbol은 character set에 의하여 computing 가능한 data의 형태로 변환될 수 있음을 보았다. 매체 자체는 단지 전달만 하는 매개체인가? 적절한 매체 자체는 정보자체를 강조할 수도 혹은 정보의 일부를 매체에서 담당할 수도 있다. 즉, 말을 할 때 (즉 매체의 모습이) 화난 목소리인지, 슬픈 목소리인지, 어린아이 목소리인지에 따라서 정보의 일부를 담당한다고 볼 수도 있다. "난 나비가 될 거야"라는 문장을 어린아이의 목소리(매체)를 통하여 전달하는 것과 근엄한 목소리의 중년남자의 목소리로 전달하는 것은 분명히 다르다. (따라서 최근에는 매체, 즉 미디어의 활용이 중요하게 부각되고 있다.) 우리가 그 언어를 어떤 목적으로 사용 혹은 computing할 것인가에 따라서 다른 data representation을 사용할 수 있다. 우리가 배운 것은 단지 symbol을 binary data로 변환하는 것을 배운 것이다.

노래는 어떠한가. 전달하려고 하는 메시지를 음악이라는 형태에 실어서 전달하는 것이다. 그것을 어떻게 computing할 수 있는 data로 representation할 것인가. 우리가 수업시간에 학습했던 audio data representation은 단지 "소리"자체만을 (정확하게 말하면 소리의 frequency)를 data로 표현한 것이다. 그 data는 우리가 noisy를 넣을 수도 혹은 어떤 특정한 frequency의 음은 자동으로 삭제되도록 하는데 사용하고자 한다면 별 문제 없을 것이다. 하지만, 음성 인식을 하여 자동으로 text문장으로 써지도록 하려고 한다면 전혀 다른 data representation을 사용하여야 할 것이다. 또는 음악을 자동으로 작곡해주게 하는 computing을 하려고 한다면 또 전혀 다른 data representation을 사용하여야 할 것이다. 만약에, 음악과 춤과 표정과 의상이 함께 사용하는 가수의 모습을 본다면 그것을 어떻게 data representation 할 것인가.

이제, 이미지 데이터를 다시 생각하여 보자. 이번에는 조금 더 구체적으로 생각하여 보자.

이미지 자체는 매체이다. 그리고 그 매체는 어떠한 정보를 나에게 전달 하려고 한다. 그리고 이미 언급하였듯이 매체 자체도 그 정보의 일부일 수가 있다.

bitmap으로 표현된 이미지는 단지 숫자의 나열이다. 그 숫자의 나열로 우리가 원하는 computing을 다 할 수 있을까. 어떤 정보를 끄집어 낼 수 있을까.



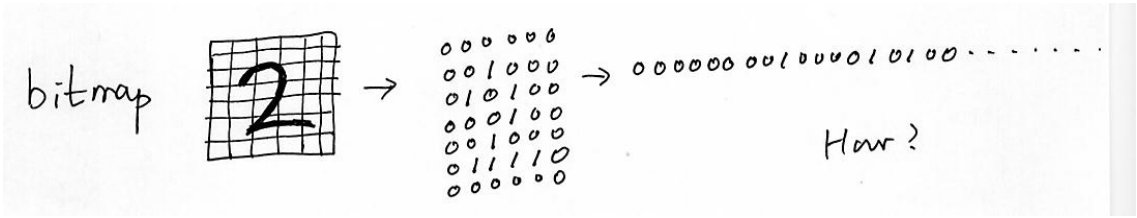
bitmap으로 바뀐 숫자들의 나열에 어떠한 computing을 가하면 위와 같은 변형된 새로운 bitmap을 만들어 낼 수 있을까. 이러한 bitmap표현은 매체자체의 data representation에 가깝다. 만약에 이러한 bitmap 데이터에서 사람의 모습만 끄집어 내어 그가 여자인지 남자인지를 구별해내는 computing을 하려고 한다면 "상당히" 어려운 작업이 될 것이다. (computing model을 찾아 내

기가 무척 힘들 것이다)

우리가 우리의 지식과 정보를 사용하여 우리가 의도하는 computing model을 만들려고 한다면, 우리는 우리의 의도에 맞게 단위정보(혹은 단위 지식)들을 효율적으로 정의할 수 있어야 한다. 예를 들어, 나는 사람들로 북적거리는 대형할인 마트인 A-Mart에 가서 카드에 물건을 잔뜩 골라서 계산대 앞으로 왔으나 20개가 넘는 계산대 마다 사람들로 길게 줄을 늘어서 있었다. **나의 "뇌"는 "어떠한 정보와 지식"을 사용하여 "어떠한 판단 모델"에 의하여 "몇 번째 계산대에 줄을 서야만 시간을 절약할 수 있겠다"라는 결정을 한다.** 어떤 정보와 지식을 사용하는지에 따라서, 그리고 어떤 컴퓨팅 모델을 사용하는 지에 따라서 나의 판단은 달라지게 되며, 나의 오후시간과 기분은 달라질 수 있다.

즉, 사람이 숫자로 쓰면, 그것을 어떠한 형태의 data로 변환하여야 하는데, 그 변환된 데이터는 내가 원하는 목적의 computing을 잘 할 수 있는 형태이어야 한다.

만약 필기체 숫자를 bitmap으로 표현하면



와 같이 변환되어 우리는 00000000100001010000010000100001111100000000 이라는 숫자에서 어떤 computing을 적용하여 이것이 2인지, 3인지를 판별하겠는가?

숫자를 쓰는 방향을 (즉 붓을 놀리는 방향을) data로 변환한다. 끊어짐도 함께 표현한다. 그러면 이 숫자에 어떠한 computing을 적용하여 이것이 2인지, 3인지를 판별 하겠는가.

요소(factor), 즉 어떠한 정류의 단위정보를 사용할 것인지를 먼저 고려하여야 한다. Bitmap은 각 pixel의 색깔만을 데이터표현의 요소로 잡았고, vector는 글자 모양의 방향을 데이터 표현의 요소로 잡았다. 먼저, 요소를 결정하고, 원 데이터를 요소에 의하여 재 표현하여야 한다. 재 표현된 요소 정보를 가지고 적절한 computing을 한다.

.끝.