한국어 전자 사전을 위한 하이퍼텍스트 네트워크에 관한 연구

이 태승*, 최 윤철
연세 대학교 전산과학과

A Study of the Hypertext Network for Korean Eletronic Dictionary

Taiseung Lee and Yoonchul Choy
Department of Computer Science, Yonsei University

요약

본 연구는 한국어 전자사전에 알맞은 하이퍼텍스트의 네트워크 구조와 전자사전의 구조에 관한 것으로 인간의 연관적 사고과정을 이용하여 사전을 구성하고자 하였다. 사용하는 사람을 계층적으로 선별하여 그에 알맞는 정보검색의 실마리를 제공하였으며 필요한 즉시 원하는 항목으로의 전환이 가능하도록 하였다. 특히 그래픽 브라우저(Graphics Browser)에 중점을 두어 사용자가 보다 편리하게 정보를 얻을 수 있도록 설계하였다.

1. 서론

기존의 국어 사전을 살펴보면 여러가지 문제점들을 안고 있음을 알 수 있다. 정작 우리들이 실용적으로 사용할 수 있는 형태의 사전이 아니라, 일본어 사전등, 기존의 사전들의 구조와 내용에 토대를 두어 구성한 듯한 인상을 주는 현재의 사전은 바람직한 한글 사전이라 할 수 없다고 하겠다. 또한 기존의 한글사전들은 한국어 정보처리 영역에서 이용하기에는 전혀 부적절하다고 하겠다. 따라서 현재까지와는 다른 형태와 내용의 사전의 개발이 시급히 요청되고 있다. 또한 한글 사전을 전자사전화 하는 작업은 앞으로의 국어학및 문학의 발전을 위해서도 꼭 필요하다고 할 수 있다.

인간은 어떤 생각이 떠오르면 그와 관련된, 또는 연상되는 생각으로 재빠르고 자유롭게 이동하며 생각하게 된다. 그러나 현재의 모든 문서나 사전과 같은 것들은 이와는 달리 선형 순차적으로 구성되어 있다. 그러므로 이들의 구조를 인간의 사고형식과 같이 연관적으로 바꾸는 작업이 필요하다. [3,5,9] 이런 의미에서 본 연구에서는 한글 전자사전을 하이퍼텍스트(Hypertext)개념을 적용하여 구성한다.

언어가 정보화 사회에서 가장 중요한 정보임은 의심할 여지가 없다. 따라서 최근에 세계 여러 나라들이 대규모 언어 정보 데이타 베이스(lexical database) 구축과 전자 사전 개발에 박차를 가하고 있다. 대표적으로 영국 옥스포드 대학의 O.E.D와 이를 더욱 발전시킨 워터루 대학의 New O.E.D가 있다. 연세 대학교 전자 사전 개발실에서도 1989년 이래 한국어 전자사전의 개발을 연구중이다. [4,17]

전자사전은 자연어 처리 기법을 이용하여 대량의 우리말 어휘 뭉치(corpus)를 정보처리 함으로써 우리말 언어 세계를 신속하고 정확하게 반영해 줄 수 있다. 여기에 하이퍼텍스트기법을 도입하여 사람의 인지 심리와 같이, 연관적 기억과 사고작용에 따른, 가장 자연스럽고 직접적이며 편리한 사전 검색 방식을 제공할 수 있도록 하는것이 본 연구의목적이다.

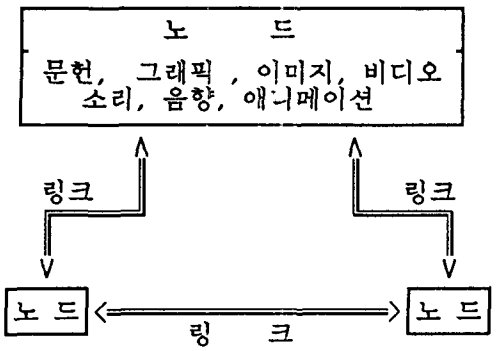
2. 시스템의 구성

우리들이 앞으로 개발해야 할 내용은 말뭉치의 수집과 그를 이용한 사전의 편찬으로 나눌 수 있다. 현재의 사전을 보면 단어중심의 사전이라 볼 수 있으며 우리는 이와는 다른 KWIC (Key Word In Context) 를 이용하여 사전을 구성하고자 한다. 따라서 어느 단어가 어떤 문헌자료에서 어떤 문맥에서 인용되었는지에 대한 정보도 역시 들어 있어야 한다. 우리가 사전을 찾는 경우는 단어의 뜻을 몰라서 찾는 경우가 대부분이다. 그러나 이것을 좀 더 정확한 의미로 말한다면 그 말이 문장중에 어떤 형식으로 쓰여지는지를 알기위해 찾는다고 볼 수 있다. 그러므로 개발하고자 하는 전자 사전에는 그 말이 인용된 자료가 등록되어 있어야 하고 그말이 쓰여진 예문들이 충분히 제시되어 있어야 한다. 그 이외에도 맞춤법, 본디말, 외래어등에 대하여 사전편찬시에 충분히 생각해야 할 것이다. [1,2,4]

1) 하이퍼텍스트(Hypertext)

가장 자유스럽고 편리한 정보 시스템은 인간의 사고 과정을 통하여 이루어 진다. 인간은 어떠한 생각이 떠오르면, 바로 그 생각과 연관되는 또는 연상되는 생각으로 직접 융통성있게 움직인다. 그러나 많은 양의 정보를 영구히 기록하거나 타인에게 전달하는 데에는 인간의 사고보다 책, 녹음기 같은 기존의 문서화 도구가 필요하다. 지금까지의 이러한 도구들은 대체로 선형 순차적 구조로 되어 있어 때때로 이용하기에 불편하고 또한 비 효율적이라는 문제점이 생기게 된다. 따라서 인간의 사고와 유사하게 필요한 정보를 얻을 수 있도록 정보를 구성, 검색할 수 있게 하이퍼텍스트를 이용해야 한다. 하이퍼텍스트는 노드와 링크의 관계를 이용하여 구성되어진 비 순차적 텍스트(text)라고 할 수 있다. [3,5,15]

현재까지 개발된 하이퍼텍스트의 문제점은 "Cognitive Overhead" 와 "Lost in Hyperspace" 라고 할 수 있다. 현재의 하이퍼텍스트는 노드(node)와 링크(link)의 구조를 이용하여 구성된다. 따라서 우리가 사용하고자 하는 한국어 사전을 하이퍼텍스트화 하자면 그에 맞는 노드구조의 설계가 필요하다. [7,11]



2) 노드(Node) 구조

모든 정보가 그렇듯이 전자사전을 구성하는 정보도 어떤 레코드(record)의 형식을 취해야 한다. 하이퍼텍스트에서의 노드는 한 화면에 나타낼 수 있는 최소의 정보를 의미한다. 그러므로 우리가 얻고자 하는 정보는 여러개의 노드로 구분되어 있으며 그 노드들은 각각 링크에 의해 연결되어 있다. 사용되는 노드의 종류는 대략 다음과 같다. [9]

■ 문(text) 노드

문으로 이루어진다. 문 자체는 문서이거나, 혹은 문서에서 제공되는 것과 같은 기본 수준의 정보를 표현하기도 한다. 문은 지식 표현의 형식에 맞게 미리 처리되지 않은 정보로서, 읽혀지도록 고안되어져 독자가 그 문으로부터 지식을 뽑아내야 한다. 따라서 사람외의 추론 시스템에는 적합한 입력은 아니다. 전자사전은 사람을 대상으로 한 시스템이므로 낱말의 정의, 용례 등을 보여 주는 정보 표현의 주 매개체로 이 노드가 사용된다.

■ 그림(picture) 노드

그림은 문 노드내에 삽입되거나 독자적인 하나의 노드가 될 수 있다. 확대 링크를 지닌 그림은 기술 서적에서 한 부품이 여러 부분으로 분해된 모습을 보여 주는 것처럼 더욱 더 상세히 보여질 수 있다. 이는 한국어 전자사전에서 인물, 동식물, 사물 등을 보여 주는 주 수단으로 사용된다.

■ 소리(sound) 노드

그림과 유사하게 소리는 문내에 삽입되던가 독자적인 노드로 존재할 수 있다. 소리는 그림에서와 같은 확대 효과는 중요하지 않지만 일반적으로 그림 노드와 비슷한 방식으로 다루어진다. 이는 낱말의 발음, 동물의 울음소리, 자연음 등을 위해 사용될 수 있다.

■ 혼합 매체(mixed-media) 노드

문, 그림, 그리고 소리의 임의 조합으로 형성된다. 많은 경우에 동일 정보가 링크된 노드의 조합에 의해서나 단일 혼합 매체 노드로써 표현될 수 있다. 이는 단순한 낱말을 설명할 때라도 한 노드 유형으로만 정보를 표현해서는 즉각적으로 이해하기 어려운 경우에 적합하다.

예를 들어 '개'를 기존 사전에서 찾아 보면 다음과 같이 정의되어 있다.

개: 개과의 짐승. 가축으로, 이리·늑대와 비슷하나 성질이 온순하고 영리함. 품종이 많음

이는 개를 전혀 모르는 사람에게 직관적으로 그 모습이나 특징을 떠오르게 하기에는 미흡하다. 대신 개의 그림과 함께 개 울음 소리를 내주는 것이 더욱 효과가 있을 것이다.

노드에 저장되어지는 정보는 실제로 이 노드가 가지고 있는 문서 정보와 여기에 연결되어지는 링크의 자료구조이다. 실제로 이런 정보를 동시에 하나의 자료구조로 표현하기는 매우 어려운 일이므로 다음과 같이 나누어 표현한다. 첫째, 노드의 문서 정보를 그대로 텍스트로 가지고 있는 데이타 필드(data feild)와 둘째, 노드와 관계되어진 링크나 그 이외의 정보를 가진 노드 필드(node feild)이다.

노드의 자료 구조는 다음과 같다.

```
Struct Node_Info {
        name_type   doc_name, node_name;
        file_type   file_name;
        struct  contents  head;              /*문서 정보*/
        struct  links  link_head;            /*링크 정보*/
        };
```

링크의 구조는 자신의 링크 Type과 목적 Filename, 그리고 다음번 또는 이전 링크에 대한 정보를 가지게 된다.

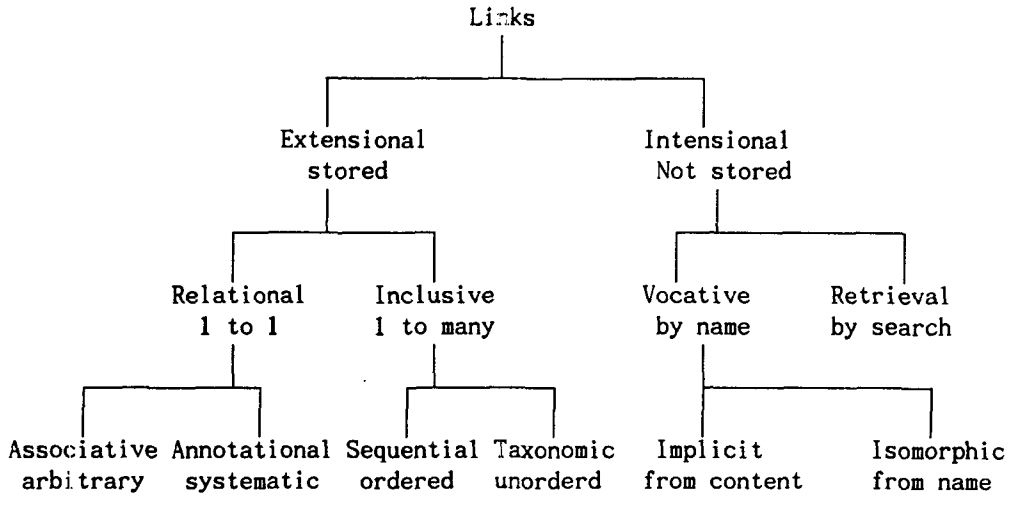
3) 레코드 구조

레코드의 내용은 노드가 지니는 문서 정보를 나타내며 따라서 하나의 단어가 가지는 여러가지 정보를 모두 표현하여야 한다. 단어가 가지는 최소한의 필드(field)는 다음과 같다.

① 표제어 ② 어원 ③ 품사 (형태적, 의미적, 기능적)
④ 활용형 ⑤ 의미 ⑥ 용례
⑦ 유사어 ⑧ 상대어
⑨ '하다'가 붙어 동사가 될수 있는지의 여부와 그 종류
⑩ '되다'가 붙어 동사가 될수 있는지의 여부와 그 종류

품사의 의미적 분류는 그 단어가 어떤 종류인가를 나타낸다. 이를테면 사과란 것은 기능적 분류로는 명사이지만 의미적 분류로는 음식, 먹을 것의 범주에 들게 된다. 활용형은 동사의 경우에 그것이 어떻게 변화 하는가를 나타낸다. 마지막 두개의 항목은 수록되어질 단어의 항목을 줄이기 위하여 만들어진 것이다. 이중에서 유사어와 상대어는 모두가 키워드(Key Word)이며 따라서 단순히 마우스(mouse)로 클리킹(clicking)함으로써 그 단어로 이동하게 된다.

4) 링크 구조
여기에 사용되어지는 링크들은 다음과 같다. [7]



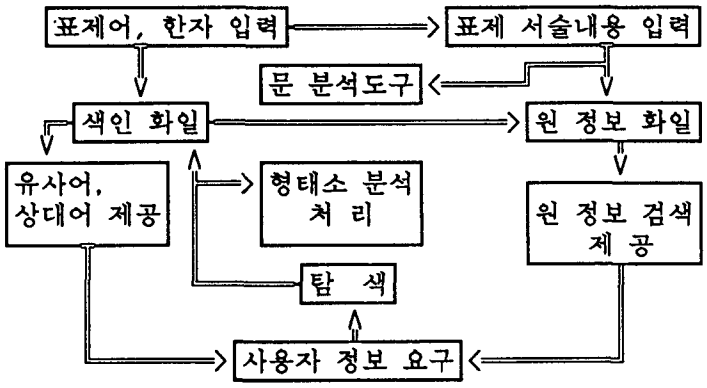
Associative 링크는 여러가지 목적으로 사용되며 보통 그 타입(type)에 따라 레이블(Label)이 붙게 된다. 이는 목적지가 어디인지 알수 없을 때, 즉 새로운 노드를 형성할 때 사용된다. Annotational 링크는 목적지가 어디인지 알 수 있는, 따라서 이미 존재하는 노드에 연결할 때 사용된다.

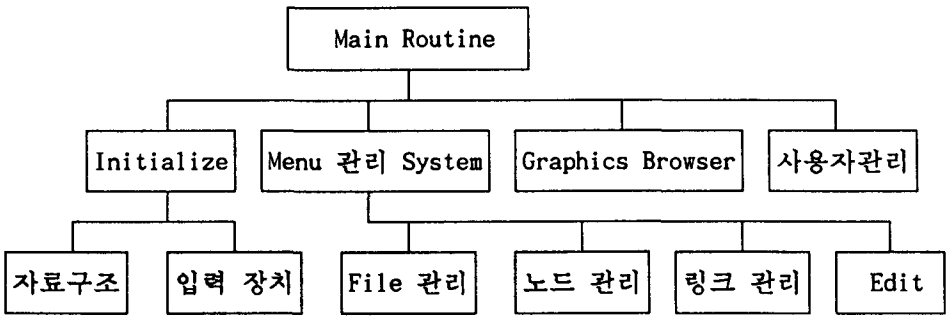
위에서 언급한 두개의 링크는 Relational 링크로 하나의 목적지를 지닌다. 사전을 구성할 때 필요한 링크는 목적지가 두개일 필요가 없으며, 만일 두개가 필요하다고 할 때에는 어느 특정 필드에 표시를 해두는 방법을 이용한다. 이 두개의 링크와는 달리 특별히 링크로 지정되지 않고 단어에 의해 찾아가게 되는 것을 Vocative 링크라고 한다. 이 링크는 다큐먼트(document)의 이름이 사용자가 이용할 수 있는 의미있는 이름일 때 쓰여진다. 그러나 한국어 사전의 경우에 있어서 모든 단어를 Vocative 링크로 처리하기에는 그 오버헤드(overhead)가 너무 크다. [7]

이외의 다른 링크들도 많이 있지만 여기에서는 위의 세가지 링크를 사용하여 전자사전을 설계하였다.

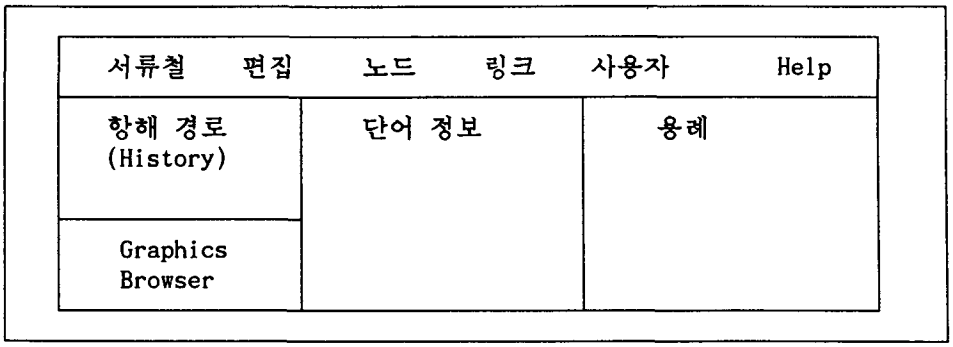
5) 화일(file)과 노드(node)와의 관계
하이퍼텍스트에서는 하나의 정보는 다큐먼트(document)라고 불리우며 이는 다시 여러개의 노드로 나뉘어 진다. 한글 사전에서 주 정보는 단어이며 단어 하나하나가 하나의 다큐먼트가 된다. 이때 각각의 다큐먼트마다 하나의 화일을 대응시킬 경우 단어의 수만큼 화일이 필요하게 된다. 이럴 경우 단어별로 구분이 되어 편한 점도 있으나 그만큼 디스크 접근(Disk Access)이 많이 필요하게 되며 따라서 찾는 시간(Searching Time)이 길어지는 결과를 초래하게 된다. 따라서 여러개의 단어를 하나의 화일에 묶어 넣는 경우를 고려해 보아야 한다. 따라서 단어를 여러 개의 뭉치로 나누는 작업이 필요하게 되며 어떤 종류의 단어를 묶을 것인가가 또한 문제가 된다. 선형 순차적인 방법으로는 가나다 식 분류가 찾기 쉬우리라 생각이 되지만 하이퍼텍스트는 비 선형이므로 그보다는 의미가 비슷하거나 활용이 비슷한 것끼리 모아 놓는것이 더 나으리라고 생각한다. 여기에 대해서는 앞으로도 많은 연구가 필요하다. [16]

6) 전반적인 시스템 구성도





7) 화면 구성도



8) 그래픽 브라우저(graphics browser)

그래픽 브라우저는 현재 화면에 나타나는 노드와 그와 관련된 노드들을 그림으로 표현해 준다. 사전에는 수십만개의 단어(node)가 존재하므로 이를 일일이 모두 표현하기는 어렵다. 따라서 어안 보기(fish-eye view)를 적용하여 중요한 몇몇 노드들만을 보여주게 한다. 또한 그래픽 브라우저에서 노드를 클리킹함으로써 그 노드로 이동이 가능하므로 사용자가 편리하게 이용할 수 있다. [3]

3. 결론

이미 외국의 많은 사전들이 전자사전의 형태로 개발되어 있으나, 하이퍼텍스트의 개념이 적용된 경우는 그리 많지 않다. OED의 예를 보더라도 하나의 키워드를 찾는 방식은 링크에 의한 것이 아니라 Text matching방식을 사용하고 있다. [17]

어느 면에서 보면 현재의 사전들은 대부분 데이타베이스(database)형식을 지닌다고 할 수 있고 OED도 일종의 데이타베이스라고 볼 수 있다. 데이타베이스에 비해 하이퍼텍스트화 된 사전은 모아둔 정보들을 어떻게 효율적으로 연결하고 사용하느냐에 따라 그 성능이 평가된다.

하이퍼텍스트와 전자사전을 결합시켜 놓은 하이퍼텍스트화된 전자사전은 다음과 같은 많은 장점을 기본적으로 내포하고 있다. [1]

- 다양한 계층 구조로 관련 정보를 한 곳으로 쉽게 모아준다.
- 비선형 문을 제공하여 Browsing을 용이하게 해준다.
- 다중 매체를 사용하여 사람의 이해를 증대시킨다.
- 화면에 보여줄 자료 크기와 디스플레이 방식 조정이 가능하다.
- 대화식 방식으로 동의어,반의어,용례의 참조로 문장 작성시 온라인상에서 자유롭고 손쉽게 잘라 붙이기(cut and paste)나 대체를 위해 사전의 효율적 사용이 가능하다.
- 문장 데이타베이스가 여러 장소와 기계 사이에 공유되어 문장을 읽고 저술하는데 상호 협력을 가능하게 한다.

그리고 제시할 수 있는 가장 이상적인 하이퍼텍스트화된 전자사전의 모형은 다양한 문, 문헌, 그림, 사진, 도표, 이미지에 관한 정보를 원시 형태 그대로 스캐너(scanner)를 통해 입력받아 전자사전 편찬기가 자동적으로 내부 정보 처리를 하여 생성할 수 있는 최적의 사전 형태로 구성해 주고 사람이 그 생성된 사전 형태로부터 가다듬는 과정을 거치고, 용례같이 완전 자동 작업이 가능한 항목은 최신의 것으로 수시로 갱신될 수 있도록 해줄 수 있고 풍부한 하이퍼미디어를 가지고 사람의 연관적 사고 과정을 그대로 지원해 줄 수 있는 전자사전 항해기(navigator)를 가진 시스템이라 할 수 있다.

4. 참고 문헌

[1] 양단희, "한국어 전자사전 원형의 설계및 구현", 연세대 대학원 석사학위 논문, 1991.

[2] 이상섭, 남기심 외, 새 한국어 사전 편찬을 위한 사전 편찬학 연구, 제1집, 제2집, 탑 출판사, 1988.

[3] 정희영, "하이퍼스페이스 상에서 효율적인 탐색을 지원하는 하이퍼텍스트 시스템", 연세대 대학원 석사학위 논문, 1990.

[4] 최윤철, 송만석, "한국어 전자사전 개발의 현황과 과제", 한국정보과학회 국어정보처리 춘계워크샵 학술 발표논문집, 1990.

[5] Jeff Conklin, · Hypertext: An introduction and survey, IEEE Computer 2, 9 (sept. 1987), pp.17-41.

[6] H. Van Duke Parunak, Hypertmedia topologies and user navigation, Hypertext'89 procedding, pp. 43-50.

[7] Steven J. DeRose, Expanding the notion of links, Hypertext'89 Proceddings, pp. 249-258.

[8] Ben Shneiderman and Greg Kearsley, Hypertext Hands-on!, Addison-Wesley, 1989.

[9] Kamran Parsave, Mark Chignell, Setrag Khoshafian and Harry Wong, Intelligent database, Wiley, 1989.

[10] Ray McAleese and Catherine Green, Hypertext: State of the Art, intelect, 1990.

[11] Carol J. Anderson and Mark D. Veljkov, Creating Interactive Multimedia: A Practical Guide, Scott, Foreman and Company, 1990.

[12] Jakob Nielson, Hypertext and Hypermedia, Academic press, 1990.

[13] Tabuchi M, Yagawa Y, Fujisawa M, Negishi A, Kojima K and Muraoka Y, Hyperbook: A Multimedia Information System That Permits Incomplete Queries, International Conferance On Multimedia Information Systems'91, pp. 3-16.

[14] Forbes J Burkowski, Textriever: A Retrieval Engine for Multimedia Databases, International Conferance On Multimedia Information Systems'91, pp. 71-76.

[15] Michael A Shepherd, Virtual Structure for Hypertext, International Conferance On Multimedia Systems'91, pp. 201-220.

[16] Ruben Caudillo and Michel Mainguenaud, An Hypertext-Like Multimedia Document Data Model, International Conferance On Multimedia Systems'91, pp. 221-242

[17] D.R. Raymond and F.W. Tompa, Hypertext and the Oxford English dictionary, Comm. of the ACM, July 1988.