단락의 의미구조에 의한 전문DB탐색의 타당성 연구

A Study on the Validity of Full-Text DB Search Based On Semantic Structure Of Paragraphs

남궁 황, 이 두 영
(중앙대학교 문헌정보학과)

Nam-Goong Whang, Lee Doo-Young
Dept. Of Library & Information, Chung Ang University

비구조화된 전문DB는 이용자가 필요로 하는 부분을 지정하여 다양하게 탐색할 수 없고, 또한 부적합 정보가 다량 탐색된다는 단점이 있다. 이러한 문제점을 해결하기 위해서 문헌의 구조화가 시도되고 있으나, 지금까지는 대부분 형태적 논리구조에 기반한 탐색기법이 연구되어 왔다. 따라서 본 논문에서는 문헌의 구조화에 형태적인 논리구조 뿐만 아니라 의미적인 논리구조도 반영될 수 있도록 단락을 객체단위로 한 의미구조 관계를 규명하고 이를 전문DB 구조에 적용하기 위한 타당성을 고찰하였다.

1. 서 론

디지털 도서관을 구축하는데 있어서 무엇보다도 중요한 것은 문헌의 디지털화 작업이라고 말할 수 있다. 문헌의 디지털화는 단순히 문헌을 기계가독형 중심으로 코드화한 죽어 있는 디지털 문헌(Dead Digital Documents)이 아니라 저자가 문헌에서 표현하고자 하는 의도, 인쇄문헌의 본질적인 특징, 문헌내용의 구조, 이용자들의 정보요구조건 등 종합적인 환경을 고려하여 형태적 논리구조와 의미적 논리구조를 적절히 결합하여 문헌에 기술된 내용을 충실히 반영하는 살아있는 디지털문헌(Live Digital Documents)이 되도록 구성하는 것이 매우 중요하다.

현재 국내에서도 디지털 도서관을 구축하기 위한 전문DB 생산 및 탐색기법에 대한 연구가 활발하게 진행되고 있다. 그러나 전문DB의 탐색과 밀접한 관련이 있는 문헌의 구조화에 서명, 목차, 장, 절 등과 같은 형태적인 논리구조를 반영하므로써 특정한 부분을 탐색할 수 있는 여건이 마련되었지만, 실제 내용이 전개되는 의미 구조를 기반으로 하는 탐색기법에 대한 연구는 미흡한 실정이다.

따라서 본 논문에서는 이점에 착안하여 문헌의 내용전개방식에 따른 의미구조의 기본단위를 단락으로 삼고, 형태적 논리구조와 결합하여 단락탐색이나 탐색된 단락을 단서로 그 관련된 부분과 상위 범주가 탐색될 수 있도록 유연성이 높은 전문DB탐색의 타당성을 고찰하고자 한다. 이를 위해 언어학에서 텍스트의 내용전개는 단락을 단위로 일정한 방식과 원칙에 입각해서 기술되고, 독자들의 텍스트에 대한 요약 및 이해능력은 단락의 전개방식이나 단락간의 의미구조 관계를 어느 정도 파악하는냐에 따라 영향을 받는다는 단락이론을 그 기반으로 삼고자 한다. 다만, 실제 시스템 설계 및 평가는 향후 연구과제로 미루고자 한다.

2. 전문DB의 구조화 : 형태적 논리구조와 의미적 논리구조

전문DB와 서지DB의 큰 차이점 중의 하나는

탐색과 동시에 전문이 제공되는가의 여부이다. 전문DB의 탐색은 원하는 전문이 즉시 제공되는 장점이 있는 반면, 서지DB는 탐색한 후 별도의 절차를 거쳐 필요로 하는 전문을 입수하게 된다. 이와 같은 전문의 장점으로 인해 최근 정보검색분야의 연구대상이 서지DB에서 전문DB로 그 초점이 이동하고 있다.

전문은 논리적인 전개순서에 따라 저자가 전달하고자 하는 모든 내용을 담고 있으며, 이들 내용들은 각각 상호 유기적인 관계를 맺고 있기 때문에 특정한 내용구조를 유지하고 있다. 예를 들면, 어떤 절의 하부에 나타난 여러 개의 단락은 각각 내용상 밀접한 관계를 가지고 있는 것이다. 이러한 전문의 구조와 특징을 고려하여 전문DB 구조화가 이루어져야 한다.

그러나 비구조화된 전문DB에서는 단지 전문내에 출현하는 모든 용어를 탐색어로 선정, 탐색연산자와 함께 결합하여 탐색할 수 있는 기능은 있으나 이용자의 요구조건에 맞는 필요한 부분을 탐색할 수 없고, 부적합 정보가 다량 탐색되므로 검색 효율성이 저하된다는 단점이 있다. 따라서 이러한 문제점을 극복하고 전문의 장점을 최대한 활용하기 위해서는 검색시스템이 전문의 내용구조를 인식하여 이를 이용자에게 제공하므로써 이용자의 정보요구와 전문의 내용을 일치시켜 최적의 적합정보가 제공될 수 있도록 내용구조를 반영한 전문DB 구축방법이 모색되어야 할 것이다.

본 논문에서는 전문의 내용구조를 정의하는 방식과 깊이의 정도에 따라 탐색결과와 이용자 만족도에 상당한 영향을 미친다는 가정하에서 내용구조를 형태적 논리구조와 의미적 논리구조로 구분하고자 한다. 형태적 논리구조는 일반적으로 문헌의 구조 분석시 많이 이용되는 표층구조로서 제목, 저자, 목차, 초록, 장, 절, 단락 등 문헌의 내용을 기술하기 위한 정형화된 틀 구조를 말하며, 의미적 논리구조는 문헌의 내용을 주제 단위별로 나누고 이들 주제간의 관계를 구조화한 것으로서 주제를 단락수준에서 파악하는 것을 의미한다. 전문DB의 구조화 과정시 이 두가지 내용구조방식를 적절히 결합할 경우 정보검색기법이 다양하게 확대되고 불필요한 정보를 최대한 억제하면서 적합한 정보를 제공하기 때문에 이용자의 노력과 시간을 절약할 수가 있을 것으로 기대된다.

3. 텍스트 의미구조에 대한 언어학적 접근

3.1 텍스트 의미구조의 형성원리

텍스트의 구조는 계층적 체계의 의미구조로 형성되며, 계층적인 구조의 내부는 논리적인 관계로 연결되어 있다. 즉 텍스트의 구조는 계층적인 관계와 논리적인 관계로 연결된 의미의 망을 형성하고 있다.

텍스트의 의미구조를 분석하기 위해서는 무엇보다도 먼저 텍스트에 대한 철저한 이해를 바탕으로 텍스트의 분석 단위(Unit)를 설정하고, 이 단위들이 연결되는 관계(Relation)를 밝혀야 한다. 특히 텍스트의 의미구조를 분석하는 단위는 텍스트의 종류 또는 분석의 목적이나 방법에 따라 명제, 범주(Category), 아이디어, 진술접점(Statement Node) 등 다양한 방식으로 수행되고 있다. 본 논문에서는 명제단위를 중심으로 해서 텍스트 의미구조의 형성관계를 살펴보고자 한다.

텍스트의 의미구조는 명제를 기본단위로 하여 명제들 간에 일정한 관계가 형성되므로써 응집적인 의미 단위체인 스키마가 형성되는 것이다. 명제는 하나의 서술(Predicate)과 하나 이상의 논항(Arguments)의 연결관계를 통해 구성되어 진다. 상위명제는 하위의 명제를 포함하며, 하위의 명제는 상위의 명제를 뒷받침해주는 상호간의 보완적인 의미관계를 유지하고 있는 것이다. 따라서 텍스트의 의미구조를 밝히는데 있어서의 핵심적인 문제는 명제들 간에 어떠한 관계가 작용하고, 어떤 의미구조로 응집되었는냐를 파악하는 것이 중요하다.

3.2 텍스트 의미구조의 체계

텍스트 의미구조는 의미분석의 규모를 어느 수준으로 삼는냐에 따라 미시구조, 거시구조, 상위구조로 나눌 수 있다.

미시구조는 텍스트 계층 구조의 최하위 단위로서 문장수준에서 의미의 구조를 밝히는 단계이며, 거시구조는 미시구조의 상위구조이면서 상대적인 개념으로서 문장이상의 단락수준에서 기술되는 의미의 흐름을 다루는 중간계층 단계이고, 상위 구조는 텍스트의 전체적인 구조체계에 해당하는 것으로 거시구조를 일반화한 것이다. 이들의 구조관계를 구체적으로 살펴보면, 미시구조는 개별적인 명제나 정보의 항목이 연결되는 방식을 대상으로 하는데 비해, 거시구조는 단락에 제시된 복잡한 명제나 개념들이 주제에 연결되는 방식에 중점을 둔다. 또한 상

위 구조는 여러 단락들이 종합되어 이루어진 전체로서 텍스트의 전반적인 구조체계를 대상으로 한다.

3.3 텍스트 의미구조와 탐색과의 관계

앞에서 살펴본 텍스트의 의미구조를 정보검색에 적용할 경우 문헌을 의미단위별로 나누어서 구조화할 수 있기 때문에 이용자는 전체 문헌을 읽지 않고 원하는 주제 내용이 기술된 부분만을 직접 탐색할 수 있을 것이다. Dillon 등은 연구자들이 문헌을 어떻게 읽고 이해하는가에 대한 문헌이용방법을 조사하여 그 결과를 전문DB에 반영해야 할 지침으로 ① 목차 ② 논문에 관한 간략정보(표제, 저자명, 초록, 인용문헌 등) ③ Browsing 기능 ④ 이용자가 원하는 부분의 인쇄 기능 등을 제시하고 있다. 이를 부분 탐색 관점에서 설명하면, 필요한 부분을 인쇄하는 기능은 이용자가 배경지식을 습득하거나 참조하기 위해서 전문을 다 읽을 필요가 없이 문헌의 중심내용 또는 원하는 부분만이 선택되기를 요구하는 것이고 Browsing 기능은 탐색된 부분을 단서로 하여 앞, 뒤 부분 혹은 절, 장 단위로 자유롭게 확대시켜 탐색할 수 있도록 하는 것이다. 이와 같이 전문DB 구축시 이용자의 부분 탐색의 요구조건을 만족시키기 위해서는 문헌의 구조화를 형태적 논리구조와 더불어 의미적 논리구조 차원에서 고려해야 할 필요성이 있다. 결국 전문DB에 문헌구조를 다양하고 깊이있게 반영하면 할수록 검색접근점과 부분탐색의 조건은 확대되어 진다.

텍스트 의미구조의 체계와 문헌의 구조화는 밀접한 관계를 가지고 있다. 형태적 논리구조는 텍스트 단위의 상위 구조에 해당되며, 의미구조는 거시구조와 미시구조 내용을 포함한다고 말할 수 있다. 그러나 텍스트 의미구조의 분석단위인 명제의 구조는 일정한 형식을 갖추고 있지 않으며, 명제를 자세히 나누는 것도 쉽지 않기 때문에 현재로서는 정보검색에 명제의 의미구조를 반영하여 활용하기는 어렵다. 이에 따라 명제 자체의 구조보다는 명제들 간의 의미연결에 의한 일정한 관계가 형성되어 의미구조로 응집되는 거시구조의 수준에서 단락의 이론을 도입, 분석하여 단락의 의미구조를 밝히고자 한다.

4. 문헌의 의미구조화를 위한 단락분석

4.1 단락의 의미단위의 기준

단락의 형식은 일반적으로 들여쓰기로 표시되는데, 단락이 의미를 갖는 정보전달단위로서의 역할을 하기 위해서는 형식과 내용이 일치되어야 한다. 단순히 들여쓰기만 하고 이에 상응하는 내용을 갖추지 못한 단락은 단락으로서의 존재가치가 없는 것이다. 이처럼 이론적인 관점에서 단락의 형식은 그 내용과 분리해서 생각할 수가 없는 것이다.

본 연구에서는 실제 많은 글들이 단락의 형식과 내용이 일치하지 않는 다는 점에서 단락의 들여쓰기 형식에 집착하지 않고, 내용적인 통합과 정보검색에 적합한 응용성을 고려하여 단락의 의미단위를 설정하고자 한다.

단락을 의미단위로 구별하는 기준은 단락간의 긴밀도와 의존도를 그 단서로 삼았다. 즉 ① 이어지는 단락의 문두에 대용어나 접속어 사용여부 ② 계속되는 단락의 첫 문장의 주어 영역에 앞 단락에서 기술된 명사어가 반복되어 나열되는 경우 ③ 내용상의 전개 방식(원인에서 결과 등) 등을 판단기준으로 하였다. 다만, 개조서(첫째, 둘째 혹은 ①, ② 등), 공식, 알고리즘, 한 문장으로된 단락 등은 상호 관련성이 긴밀한 단락에 포함시켰으며, 단일 단락일지라도 완결도가 강한 단락일 경우에는 하나의 의미단위로 간주하였다.

4.2 문맥상 역할에 의한 단락구분

문맥적인 측면에서 단락의 구분은 문헌내에서 어떤 목적과 역할을 수행하고, 그 중요성은 어느 정도인가에 따라 판단할 수 있다. 학자들의 주장에 의하면 단락의 목적 및 역할에 따라 도입단락, 주요단락, 결말단락, 보충단락, 회화단락, 전환단락, 종속단락, 특수단락, 강조단락, 부연단락, 종결단락, 연결단락 등 다양하게 구분하고 있다. 그러나 이들 가운데 학자들이 공통적으로 주장하는 단락의 의미단위 요소를 추출하면, 도입단락, 주요단락, 보충단락, 결말단락으로 나눌 수 있다.

한편의 논문은 여러 주제들이 엮어 작성되는데, 여기에서 상위 구조를 서론, 본론, 결론 부분으로 나누는 것 처럼 한 주제에 대해서도 단락의 의미단위를 기준으로 하여 상부구조를 구분할 수 있다. 예컨데, 도입단락은 서론부분에, 보충단락을 포함하는 주요단락은 본론부분에, 결말단락은 결론부분에 해당된다고 할 수있다.

도입단락은 글의 첫 시작을 다루는 부분으로서 이어지는 단락에서 다룰 문제제기와 내용의 윤곽만을 소개하고 구체적인 내용은 언급하지 않는다. 주요단락은 주제의 세부내용과 이에 대한 보충적인 내용을 기술한 부분이고, 결말단락은 글의 마지막 마무리를 짓는 단락으로 앞단락에서 기술된 내용을 간결하게 요약하는 부분이다. 도입단락과 결말단락은 글의 전개상황에 따라 생략이 가능하다.

4.3 주제전개방식에 의한 단락의 속성추출

주요단락은 실질적으로 텍스트의 중심을 이루는 중요한 의미단위에 해당되며, 이 단락은 주제를 전개하는데 있어서 몇개의 대,소 논점으로 나누어지고, 이들은 상호간에 긴밀성을 유지하면서 주제 내용을 뒷받침해 주고 있다. 따라서 단락을 정보검색에 적용하기 위해서는 문맥상 어떤 역할을 하느냐와 함께 그 단락이 어떠한 내용전개방식을 따르고 있느냐를 파악하는 것이 아주 중요하다. 왜냐하면, 정보를 탐색하는 이용자 입장에서는 자신이 찾고자 하는 정보의 내용이 전체적인 것일 수도 있지만 어떤 개념의 정의, 특정대상의 원인과 결과관계, 예시 등과 같은 부분적인 단락단위의 내용일 경우도 있다. 이러한 관점에서 주제내용의 핵심 기술부인 주요단락을 내용전개 방식별로 구분하여 이에 해당되는 의미를 갖는 속성을 추출할 필요성이 제기된다.

학자들의 주장에 따르면, 주요단락을 주제전개 내용에 따라 정의, 한정, 상술, 소거, 반복, 분류, 구분, 분석, 과정, 예시, 비교와 대조, 원인과 결과, 묘사, 서사, 지정, 설득, 인용, 추론, 열거 및 개괄 단락 등으로 구분하고 있다. 그러나 이러한 단락의 속성은 학자에 따라 그 종류와 개념 간에 다소 차이를 보이고 있는데, 이들에 대한 내용을 분석하여 유사한 개념은 통합하거나 새로운 용어로 대체하는 방식으로 단락의 의미를 구조화하기 위해 필요한 속성을 추출하면, 정의, 분류와 구분, 예시, 비교와 대조, 원인과 결과, 과정, 분석 및 평가, 인용, 주장, 설명적 묘사, 일반단락 등으로 정리할 수 있다. 그리고 각 속성에 대한 내용범위를 상,하위개념, 관련 및 유사개념 등으로 확대하고자 한다. 예컨데 종류는 분류와 구분에, 절차는 과정에 포함시킬 수 있는 것 처럼 각 속성의 의미를 광의로 해석하여 단락의 의미구조화에 적용할 수 있다.

또한 단락내의 주제문장의 위치에 따라 두괄식, 중괄식, 미괄식, 양괄식으로 구분할 수 있는데, 일반적으로 우리나라 글에서는 두괄식과 미괄식 단락이 주류를 이루고 있다. 이를 색인작성시 활용하므로써 검색의 효율성을 증대시킬 수 있다.

5. 결론 및 향후과제

이상 살펴 본 바와 같이 문헌의 의미적 논리구조화를 위하여 단락을 문맥상 역할과 주제내용전개방식에 따라 의미를 갖는 속성단위로 추출하고자 하였다.

전문DB의 구조화에 단락을 기반으로 한 의미적인 구조를 반영하므로써 이용자의 정보요구조건과 실제 문헌에 기술된 내용을 일치시켜 부분 탐색을 수행할 수 있기 때문에 불필요한 정보를 최대한 줄이고 최적의 적합정보를 제공할 수 있을 것으로 기대된다. 또한 형태적 논리구조와 의미적 논리구조를 적절히 결합하여 다양하게 확대된 탐색기법을 활용할 수가 있을 것이다.

앞으로 실제 검색 시스템에 의미적 논리구조를 적용해보고 이에 대한 성능평가가 이루어져야 할 것이다.

참고문헌

1. 김성혁, 디지털도서관 구축을 위한 문헌의 구조화와 디지털화에 관한 연구, 기술정보 학술대회 논문집, 국방과학연구소, 1997, PP.3-20
2. 野末道子, 段落を對象とした日本語全文データベースの檢索,Library and Information Science, No. 31, 1993, PP. 79 - 93
3. Allen, B., Text Structures and The User-Intermediary Interaction, RQ, Vol. 47, No. 4, 1988, PP. 535 - 541
4. Campbell, Martha E., Focus : From Paragraph To Essay, Prentice Hall, 1996
5. Dillon, A., Richardson, J., Mcknight, C., Towords the Development of a Full-Text. Searchable Database : Implication From a Study of Journal Usage, Britich Journal of Academic Librarianship, Vol. 3, No. 1, 1988, PP. 37 - 48
6. Macleod, I.A., Storage and Retrieval of Structured Documents, IPM, Vol.26, No.2, 1990, PP. 197 - 208