

2019

Information Retrieval Modelling

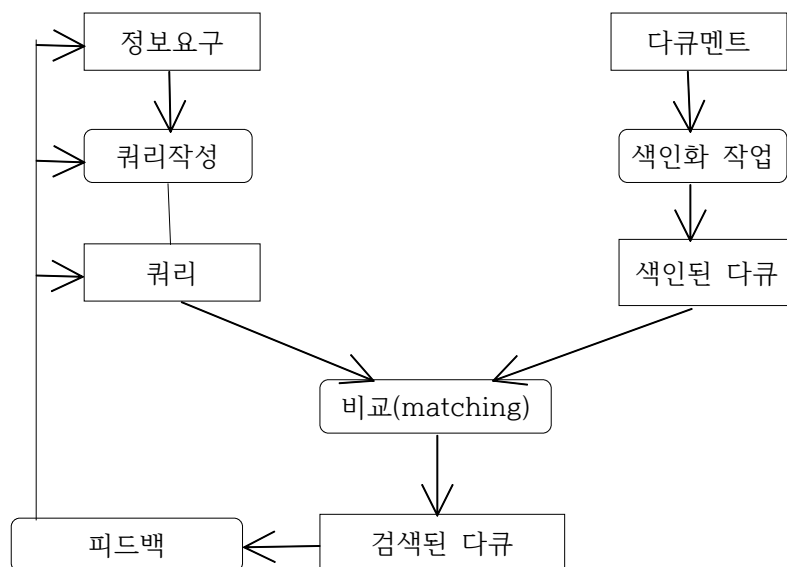
I. OVERVIEW

정보검색시스템이란 텍스트나 멀티미디어로 되어 있는 다큐에 있는 정보를 저장하고 관리하는 소프트웨어 프로그램이다. 이 시스템은 이용자들 도와 그들이 필요한 정보를 찾도록 한다. 그러나 명확하게 말해서(explicitly) 정보를 반환하거나 질문에 답하지는 않는다. 대신에, 이 시스템은 찾는 정보가 있을 있는 다큐의 존재와 위치에 대해 알려준다.

이용자의 정보요구를 만족시켜줄 수 있는 다큐를 적합한(relevant) 다큐라 부른다. 완벽한 검색시스템은 단지 적합한 다큐만을 검색할 것이며, 부적합 다큐는 결코 검색하지 않을 것이다. 그러나 이런 완벽한 시스템은 존재하지 않으며, 앞으로도 존재하지 않을 것이다. 왜냐하면, 탐색문(search statements)이 결코 완벽하지 않기 때문이며, 또한 적합성이란 이용자의 주관에 달려있기 때문이다: 어떤 이용자는 결과에 만족하지만 다른 이용자는 그렇지 않을 수 있다.

정보검색시스템이 지원해야 하는 3가지 기본적 절차:

- 1) 다큐 내용의 표현
- 2) 이용자 정보요구의 표현
- 3) 두 가지 표현의 비교.



II. Subject indexing

From Wikipedia, the free encyclopedia

주제색인(subject indexing)은 대상 다큐가 무엇에 대한 것인지, 그것의 내용이 무엇인지, 또는 그 다큐의 발견가능성(findability)을 높이기 위하여 다큐를 색인어나 기타 심볼로 기술하거나 분류하는 행위를 말한다. 다른 말로 해서, 이것은 다큐의 주제를 밝혀 기술하는 것이다. 색인은 3가지의 독립적 수준에 따라 구축된다: 1) 책과 같은 다큐 속의 용어들, 2) 도서관 장서 속의 대상물들, 3) 지식을 다루는 분야에 있는 다큐(책과 학술기사)들.

주제색인은 특별한 주제에 대한 다큐를 검색하여 특히 서지 데이터베이스를 만들기 위하여 정보검색에서 사용된다. 학술 색인 서비스의 예는 Zentralblatt MATH, Chemical Abstracts 그리고 PubMed 등이다. 색인어들은 대부분이 전문가에 의해 할당되었지만 저자 키워드 또한 일반적으로 사용되고 있다.

색인과정은 다큐의 주제분석에서 시작한다. 색인자는 그 다음에 다큐에서 직접적으로 단어들을 발췌하거나 통제어휘집(controlled vocabulary)에서 단어들을 할당하여 주제를 정확하게 나타내는 용어들을 구별하여야 한다. 그런 다음, 색인어들은 체계적인 순서로 정렬되어야 한다.

색인자는 얼마나 많은 용어를 포함시킬 것인지 그리고 그 용어들이 얼마나 전문성을 갖는지를 결정하여야 한다. 이러한 내용들을 색인의 깊이(depth of indexing)라 한다.

1. Subject analysis

색인의 첫단계는 다큐의 주제사(subject matter)를 결정하는 것이다. 수작업 색인에서, 색인자는 “이 다큐가 특별한 결과, 조건 또는 현상을 다루고 있는가?” 와 같은 질문의 해답과 관련된 주제사를 결정해야 할 것이다. 이러한 분석은 색인자의 지식과 경험에 영향을 받으므로, 내용분석에는 두 명의 색인자가 서로 다른 색인어를 추출할 수 있도록 참여하여야 한다. 이것은 검색의 성공에도 영향을 끼칠 수 있다.

1.1 Automatic vs. manual subject analysis

자동색인은 주제 카테고리를 할당하기 위하여 단어 패턴의 빈도수를 분석한 다음, 다른 다큐와 비교하는 집합과정으로 진행된다. 이것은 색인 자료에 대한 어떠한 이해력도 필요로 하지 않는다. 그러므로 더욱 더 일정한(uniform) 색인을 만들도록 하지만 참된 의미의 해석을 희생시키기도 한다. 컴퓨터 프로그램은 서술문의 의미를 이해하지 못하므로 적절한 용어를 할당하지 못하거나 부정확하게 할당한다. 깊이 있게 전문을 분석하는 것은 비용도 비싸고 시간도 소모적이지만, 인간 색인자는 서명, 초록, 요약, 결론과 같은 다큐의 특정 부분에 관심을 집중시킬 수 있다.

자동화 시스템은 시간제한이 없으므로, 다큐 전체를 분석할 수 있다. 그러나 또한 다큐의 특별한 부분만을 취급(direct)할 수 있는 option도 가지고 있다.

2. Term selection

색인의 두 번째 단계는 분석된 주제를 색인어 세트(translation)시키는 것이다. 여기에는 다큐에서 발췌하기 또는 통제어휘집에서 할당받기 등이 포함된다. 전문탐색 능력을 갖춘 많은 사람들은 정보탐색 시에 자신의 기술에 의존하면서, 전문탐색이 높은 인기를 끌고 있

다. 주제색인과 그것의 전문가, 전문색인가, 목록가, 그리고 사서는 정보 조직과 검색에 중요한 위치를 차지하고 있다. 이들 전문가들은 통제어휘를 이해하고 있으며 전문탐색으로 찾을 수 없는 정보를 찾을 수 있는 능력을 가지고 있다. 주제색인을 만들기 위한 전문가의 분석 비용은 전문탐색용 자료를 만드는데 필요한 하드웨어, 소프트웨어, 그리고 인건비에 비교될 수 없다. 모든 이용자가 다큐를 해제할 수 있는 새로운 웹 어플이 등장함으로써, social tagging 인 웹에서 크게 인기를 끌고 있다.

색인의 한 어플인 book index은 정보혁명기에도 비교적 변하지 않고 존재하고 있다.

2.1 Extraction/Derived indexing

발췌색인은 다큐에서 직접 단어를 추출하여 만든다. 이것은 자연어를 사용하며 단어빈도 계산을 사용하는, 또는 색인으로 사용하기 위하여 사전에 빈도수의 임계치를 결정해 놓은 자동화 기법에서 많이 사용하고 있다. the, and와 같은 일반 단어를 포함하고 있는 stop-list에 있는 stop words는 색인어에서 배제될 것이다. 자동발췌색인에서는 어귀가 아니라 단일 단어만을 색인함으로써 용어들의 의미가 손실될 수도 있다. 비록 상례적으로 발생하는 어귀의 발췌가 가능하다 하더라도, 만일 key 개념이 어귀 속에 일관성 없이 사용된다면 이것은 더욱 어려워진다. 자동발췌색인에서는 또다른 문제가 있는데, 일반 단어를 제거하기 위하여 stop-list를 사용할 때조차도, 몇몇 빈도가 높은 단어들은 다큐를 차별화하는데 쓸모가 없을 수 있다. 예를 들어, glucose 단어는 diabetes(당뇨병)와 관련된 다큐에서는 빈도수가 높을 수 있다. 그러므로 이러한 용어의 사용은 관련 데이터베이스에 있는 대부분 또는 모든 다큐를 결과로 얻을 수 있다.

탐색 시에, 용어들이 결합되는 후조합색인(post-coordinate indexing)은 이러한 결과를 줄일 수 있지만, 그 부담(onus)은 정보전문가의 몫이 아니라 올바른 용어들을 링크시켜야 하는 탐색자의 몫이 된다. 추가로 저빈도 용어는 매우 중요할 수 있다. 예를 들어, 의학분야에서 저빈도로 사용되는 신약은 그 주제분야에서 매우 중요한 것이다.

자동화 기법에서 빈도가 적은 용어는 포함시키고 일반 단어는 배제시키는 한 가지 방법은 relative frequency approach 이다. 이것은 다큐에 있는 단어의 빈도를 데이터베이스 전체에 있는 빈도와 비교하는 방법이다. 다시 말해서, 데이터베이스에서의 기대치보다 다큐에 더 많이 나타나는 용어는 색인으로 사용될 수 있으며, 양쪽 모두 대등하게 나타나는 용어는 배제시킨다. 자동화 발췌가 갖고 있는 또 다른 문제는 어떤 개념이 논의는 되고 있으나 색인용 키워드로 텍스트에서 판단될 수 없을 때, 그것은 인정받지 못한다는 것이다.

2.2 Assignment indexing

할당색인은 통제어휘에서 색인어를 취한다. 이것은 우선순위어(PT: preferred term)를 색인할 수 있으므로, 동의어를 통제하는데 장점이 있다. 이음동의어와 연관어(RT)는 이용자를 우선순위어로 접근시킨다. 이것은 이용자는 저자가 사용한 특별한 용어와 상관없이 학술기사를 찾을 수 있다는 것과 모든 가능성 있는 동의어에 대해 알아야 한다는 부담을 덜어준다는 의미이다. 이것은 또한 a qualifying term를 포함시킴으로써 동형의의어로 인한 혼란을 방지할 수 있다. 3번째 장점은 계층적으로나 집합적으로 연관어의 링킹을 허용한다는 것이다. 예를 들어, oral medication용 색인 엔트리에는 그것의 계층과 동일한 수준에 있는 RT로 다른 oral medications를 리스트할 수 있으며, 또한 treatment와 같은 광의어(BT)에 링크될 수 있다. 할당색인은 서로 다른 색인자들이 선택한 용어를 통제할 수 있으므로 색인자간의 일관

성을 개선시킬 수 있고, 수작업 색인에서 사용된다. 통제어휘라도 두 색인자가 계속해서 주제를 서로 다르게 해석한다면 이러한 모순을 완전하게 제거하진 못한다.

3. Index presentation

색인의 최종 단계는 체계적인 순서로 entries를 제공하는 것이다. 이것에는 linking entries도 포함된다. 전조합색인에서, 색인자는 이용자의 탐색식을 고려하여 엔트리에 연결된 용어들의 순서를 결정한다. 후조합색인에서, 엔트리들은 한 개씩 제공되며 이용자는 탐색을 통해 그 엔트리에 연결될 수 있다. 후조합은 전조합에 비해 정확성의 손실을 가져온다.

4. Depth of Indexing

색인가는 어떤 엔트리를 포함시키고, 색인에 얼마나 많은 엔트리를 통합시켜야 하는가를 결정해야 한다. 색인의 깊이란 망라성(exhaustivity)과 전문성(specificity)을 참고하는 색인과 정의 완벽성을 말한다.

4.1 Exhaustivity

망라적 색인은 모든 잠재적 색인어를 리스트하고 있는 색인이다. 망라성이 클수록 재현률이 더욱 높아지며, 모든 적절한 학술기사가 검색될 가능성도 높아진다. 그러나 이것은 정확률의 희생을 유발한다. 이것이 의미하는 것은 이용자는 더 많은 부적합한 다큐, 또는 단지 주제의 깊이 낮은 다큐만을 검색할 수도 있다는 것이다. 수작업 시스템에서 망라성의 수준이 높으면 높을수록 인력이 더 많이 투입되어야 하므로 비용도 더욱 많아진다. 자동화 시스템에서 소비되는 추가 시간은 다소 중요할 수도 있다. 이와 반대편에 있는 선택 색인(selective index)에서 이것과 관련된 중요한 요소들을 다루고 있다. 색인가가 충분한 용어를 포함시키지 않는다면 선택 색인에서 재현률은 축소될 것이며, 매우 적합한 학술기사가 무시될 수도 있다. 그러므로 색인가는 균형을 잡도록 애써야 하며 다큐를 어떻게 이용할 것인가를 고려해야 한다. 이들은 또한 시간과 비용의 관계도 고려해야 한다.

4.2 Specificity

전문성이란 색인어와 주제가 얼마나 밀접한가를 나타낸다. 만일 색인자가 다큐의 개념에 parallel descriptors를 사용하여 그 개념을 정확하게 반영한다면, 그 색인을 전문적이라고 부른다. 전문성은 망라성과 더불어 증가하는 경향이 있다. 즉, 여러분이 더 많은 용어를 포함시키면 시킬수록, 그 용어들은 더욱 더 협의적이 된다.

5. Indexing theory

5.1 Rationalist theories of indexing (such as Ranganathan's theory)

주제는 기본적인 카테고리로부터 논리적으로 구축된다는 이론이다. 주제분석의 기본적인 방법은 기본적인 카테고리의 세트를 독립(analysis)시킨 다음, 어떤 규칙에 따라 이러한 카테고리들을 결합(synthesis)시켜서 특정한 다큐의 주제를 구축하는 "analytic-synthetic" 방식이

다.

5.2 Empiricist theories of indexing

특히 수리통계 기법을 사용하여, 속성을 근거로 유사한 다큐를 선택한다는 이론이다.

5.3 Historicist and hermeneutical theories of indexing

이 것의 주장은 특정 다큐의 주제는 특정 담론이나 도메인과 관계를 갖는데, 그 이유는 색인이란 특별한 담론이나 도메인의 요구를 반영해야하기 때문이라는 것이다. hermeneutics(해석학)에 따르면, 다큐는 항상 특별한 지평(horizon)으로부터 기록되고 해석된다는 것이다. 지식조직 시스템과 그러한 시스템을 탐색하는 모든 이용자도 이것과 똑같다. 그러한 시스템에 입력되는 질문은 특별한 지평으로부터 입력된다. 모든 이러한 지평은 다소 공감하거나 또는 논쟁적일 수 있다. 다큐를 색인하는 것은 다양한 지평에 대해 지식을 갖춤으로써 “적절한” 다큐의 검색에 도움이 되어야 한다.

5.4 Pragmatic and critical theories of indexing

이것은 주제란 특별한 담론과 상관이 있다는 역사학자의 견해에 동의하지만, 주제분석은 특정한 목표와 값에 도움이 되어야 하며 여러 가지 방법으로 색인의 영향력(consequences)을 고려해야한다는 것을 강조하는 이론이다. 이 이론에서는 색인이란 중립적일 수 없으며 중립적 방법으로 색인하는 것은 목표가 잘못된 것이고 주장하고 있다. 색인은 행동(act)이다. (그러므로 컴퓨터 의존형 색인은 프로그래머의 의도에 따라 행동한다). 행동은 인간 목표를 지원한다. 도서관 정보 서비스들도 인간 목표를 지원 한다. 왜냐하면 그것들의 색인이 가능한 이러한 목표를 지원하는 방식으로 작성되었기 때문이다.