자동색인의 이론과 실제

Theory and Practice of Automatic Indexing

윤 구 호 (Koo-Ho Yoon)*

〈목 차〉

초 록

본고는 자동 발췌색인법과 할당색인법의 기법과 문제점, 인공지능에 의한 색인전문가시스템 및 인터넷에서 사용되는 주요 색인작성과 문제점 등을 분석하고, 아울러 색인작성과 연관된 하이퍼텍스트 링크의 자동설정 방법도 간략히 살펴보았다. 현대의 텍스트처리에서 사용되는 대부분의 기법은 특별히 새로운 것이 아니고 아마도 초보적 형태로 Luhn을 비롯한 많은 연구자들에 의해 이미 30여년 전에 사용된 것이지만 보다 훌륭한 결과가 달성될 수 있는 것은 대량의 전자적 텍스트가 이용가능하고 또한 이러한 텍스트 처리를 효율적으로 수행할 수 있는 컴퓨터의 능력 때문이다.

Abstract

This paper deals with the methods as well as the problems associated with automatic extraction indexing and assignment indexing, expert systems for indexing, and major approaches currently used to index the Internet resources. It also briefly reviews basic methods for establishing hypertext/hypermedia links automatically.

The methods used in much of text processing today are not particularly new. Most of them were used, perhaps in a more rudimentary form, 30 or more years ago by Luhn and many other investigators. Better results can be achieved today because much greater bodies of electronic text are now available and the power of present-day computers allows the processing of such text with reasonable efficiency.

* 계명대학교 문헌정보학과 교수

Ⅰ. 서 론

자동색인의 토대를 마련한 기초적 이론은 Zipf의[1] 연구에서 그 효시를 찾아볼 수 있으며, 특히 단어빈도수에 입각한 자동색인법은 Luhn과[2] Baxendale의[3] 연구에서 비롯되어 많은 연구와 실험이 계속되었다.

자동색인은 문헌에 출현한 단어들과 이들 상호간의 관계가 문헌의 내용개념을 나타낼 수 있다는 가정에 근거를 두고 있다. 자동색인은 문헌의 텍스트를 컴퓨터가 특수한 분석기법으로 분석한 후 문헌의 내용(주제)을 나타낼 수 있는 단어나 문구를 추출하여 작성한 색인으로, 분석대상이 되는 텍스트는 문헌의 전문이나 초록 또는 표제가 된다.

자동색인의 기본원리는 문헌을 구성하는 단어들을 일정한 기준에 의거하여 주제어와 비주제어를 구분하고, 주제어로 평가된 단어로부터 색인어를 선정하는 것이다. 자동색인의 장점으로는 보다 일관성있는 색인작업이 가능하며, 색인작성의 비용이 저렴하고, 다양한 탐색어를 제공할 수 있다는 점 등이다. Salton은[4] 자신을 포함한 여러사람의 자동색인 실험결과를 분석·평가한 결과 간단한 자동색인기법들은 처리속도가 빠르고, 비용이 적게 들며, 검색효율의 재현율과 정도율에 있어서 수작업 색인의 결과와 최소한 동등하다는 결론을 내렸다.

정보량의 급증현상, 인터넷의 등장 및 디지털도서관의 구축 등 정보서비스환경의 변화는 신속하고 효율적인 정보검색을 위한 자동색인의 필요성과 중요성을 강조함으로써 자동색인의 다양한 기법이 연구 개발되어 운영되고 있다. 본고는 현재까지 연구개발된 자동색인법 중에서 괄목할만한 기법, 인공지능에 의한 색인전문가시스템 및 인터넷상의 주요 색인작성과 문제점 등을 분석하고, 아울러 색인작성과 관련된 하이퍼텍스트/미디어의 링크방법을 살펴봄으로써 이용자에게 보다 효과적인 정보검색을 위한 자동색인의 다양한 기법을 알게하고 또한 이 분야의 연구에 다소나마 도움을 주고자 함에 그 목적이 있다.

---

1) Zipf, G. K, *Human Behavior and the Principle of Least Effort.* Cambridge : Addison-Wesley, 1949.
2) Luhn, H. P, "A statistical approach to mechanized encoding and searching of literary information," *IBM Journal of Research and Development,* Vol. 1(1957), p. 309-317.
3) Baxendale, P. B, "Machine-made index for technical literature-an experiment," *IBM Journal of Research and Development,* Vol. 2(1958), p. 354-361.
4) Salton, G. and McGill, M. J. *Introduction to Modern Information Retrieval.* New York : McGraw Hill, 1983. pp. 99-110.

Ⅱ. 자동색인의 기법과 문제점

색인법에는 기본적으로 할당색인법(assignment indexing)과 발췌색인법(extraction indexing)의 두 종류가 있다. 전자는 색인작성자가 문헌의 내용을 분석하여 추출한 주요 개념들을 어느 정도 통제된 색인어휘집(일반적으로 통제어휘집이라고 함)에서 색인어(디스크립터)를 선택하여 부여하는 방법으로 작성된 색인으로서, 대부분 인위적 색인작성이라고 볼 수 있는 전통적인 통제언어색인이 이에 해당된다. 반면에 후자는 문헌의 내용을 분석한후 추출된 주요개념들을 나타내는 단어나 문구들이 저자가 사용한 자연언어 그대로 색인어로 채택되는 방법으로 작성된 색인으로서, 유니텀색인이나 대부분의 자동색인이 이에 해당된다. 그러나 오늘날의 자동색인법은 전통적인 발췌색인법을 포함하여 보다 효율적인 정보검색을 위한 할당색인법에 관한 연구와 실험이 꾸준히 지속되고 있다.

1. 발췌색인법(Extraction Indexing)

디지털도서관 환경에서는 텍스트가 전자적 형태로 축적되기 때문에 컴퓨터가 색인어의 출현빈도수, 위치 및 문맥의 전후관계의 기준 등을 이용하여 발췌색인을 작성할 수 있다. 간단한 기법은 텍스트에 있는 단어(불용어 제외)들의 출현빈도를 계수하여 순위화한 후 빈도수가 높은 단어들을 일단 색인어로 선정하는 것이다. 이를 위한 분리점(cutoff point)은 단어들의 절대빈도수, 텍스트길이에 관련된 빈도수, 또는 어떤 기준치 이상의 빈도수에 따라 설정될 수 있다.

약간 복잡한 기법은 텍스트에서 자주 출현하는 중요 문구를 발췌하여 문헌을 단어와 문구의 결합으로서 표현할 수 있는데, 이때에는 문구선정을 위한 빈도수기준이 중요한 단어선정의 기준보다 덜 엄격해야 한다. 또한 단어의 어근에 의해 관련된 모든 단어를 검색하는 용어절단기법과 단어, 문구, 또는 어근들의 출현빈도수를 반영하는 가중치를 부여하여 색인어의 중요도를 산출하는 가중치기법도 사용될 수 있다.

빈도수기준은 다른 기준에 의해 보완될 수 있다. 예를들면, Baxendale은[5] 문헌내에서의 문장의 위치가 문헌의 중요내용과 밀접한 관계가 있다는 사실에 주목하고, 모든 문단의 첫번째 문장과 마지막 문장에서 출현하는 단어를 색인어로 선택하는 방법을 제안하였다. 이는 실제로 그녀가 실험집단으로 선정한 200개 문단 중에서 85%가 첫 번째 문장에, 그리고 7%가 마지막 문장에 "주제문장"(topic sentence)이 있음을 발견한 연구에 근거하고 있다. 초기의 자동색인

5) Baxendale, P. B, op. cit.

법에서는 텍스트의 "중요정보"(information rich)부분을 식별하기 위한 여러 가지 방법들이 제안되거나 실험되었다. 예컨대, 전치사구, "결론"(conclusion)이나 "요약"(summary)과 같은 단서어에 따라오는 텍스트, 그리고 최초로 나타나는 명사를 포함하는 텍스트 등과 같은 요소를 탐색하는 기법들이 연구되고 실험되었다.

단순히 단어나 문구의 빈도수를 사용하여 용어를 선정하는 방법의 한가지 분명한 단점은, 불용어리스트를 사용한 후 조차도, 문헌에서 자주 출현하는 어떤 단어들은 DB내에 있는 다른 문헌들로부터 이 문헌을 구분할 수 있는 좋은 식별어가 될 수 없다는 점이다. 왜냐하면 이 단어들은 또한 DB내에서도 전체적으로 자주 출현하기 때문이다. 분명한 예를 하나 들면, 'library'와 'information'이라는 단어들은 문헌정보학 집서내에 있는 개별 아이템들의 매우 좋은 식별어가 될 수 없는 것이다. 그러므로 어떤 특정문헌에서 'asbestos'라는 단어는 4번 출현하는 반면에 'library'라는 단어는 12번 출현함에도 불구하고 'asbestos'가 훨씬 좋은 식별어가 될 수 있다. 왜냐하면 그것이 문헌정보학 문헌에서 드물게 출현하는 용어이기 때문이다.

단어출현빈도는 텍스트처리에서 염려할 유일한 빈도는 아니다. DB내에서의 한 단어의 전체적 출현빈도는 더욱 중요하다. 즉, 가장 좋은 식별어가 되는 단어들은 집서에서 예기치 않은 희귀한 단어들이다. 예를 들면, 도서관학에서의 'asbestos'나 asbestos 회사의 DB내에 있는 'library'같은 용어들이다. 실제로 완전한 텍스트 DB에서의 특정단어의 출현빈도를 계산할 필요는 없고 다만 텍스트탐색을 위해 사용되는 도치파일내에서의 그 단어의 출현빈도는 필요하다(즉, 파일내에 있는 모든 단어들의 출현빈도수에 관련된 특정단어의 출현빈도수를 의미함).

문헌에서의 특정단어의 절대빈도수보다는 오히려 용어선정을 위해 상대빈도수가 사용될 수 있다.[6] 이 방법에서는 단어나 문구가 DB내의 전체적 빈도율보다 어떤 특정문헌에서 더 많이 출현했을때 선정된다. 이 방법은 절대빈도수 방법보다는 약간 더 어렵다. 왜냐하면 DB내 단어들의 총빈도수에 관련한 각 단어의 빈도수의 유지와 또한 각 단어의 빈도율과 특정 문헌에서의 특정단어의 빈도율과의 비교를 요구하기 때문이다.

용어발췌의 기준으로는 절대빈도수, 상대빈도수, 또는 양자의 조합은 물론 위치 또는 구문적 기준 등이 있다.[7] 만일 용어선정을 위해 상대빈도수가 사용된다면 물론 불용어리스트는 실제로 필요가 없다. 전치사, 접속사 및 관사들은 개개의 아이템에서 자주 출현할 것이나 그들은 또한 DB에서도 자주 출현할 것이므로 도서관학에 있어서 '도서관'이란 용어처럼 실명사지만 상례적으로 출현하는 용어들과 함께 제거될 것이다.

용어들은 (어떤 유형의)사전에 축적된 "용인된"(acceptable) 용어들과 일치될 때 텍스트로부터 또한 발췌될 수 있다. 이는 1970년대에 미국의 DDC(Defense Documentation Center)에서

---

6) Oswald, V. A., Jr. et al, *Automatin Indexing and Abstracting of the Contents of Documents*, Los Angeles : Planning Research Corporation, 1959. RADC-TR-59-208.

7) Salton, G. and McGill, M. J, 1983. op. cit.

수행된 MAI(machine-aided indexing)에 관한 중요한 연구에 근거를 두고 있다.[8] 본질적으로 표제와 초록에 출현한 단어열(word strings)이 자연어데이터베이스(NLDB)와 대조되어 일치된 단어열은 색인용어 후보가 된다. 즉, 이들은 DDC 시소러스로부터 디스크립터로 할당될 수 있다. Klingbiel과 Rinker는[9] MAI와 수작업색인의 결과를 비교하였는데, 3개의 사례연구결과 편집되지 않은 MAI는 재현율 수준에서는 수작업색인에서 달성된 수준에 필적할만 하였으며, 정도율 수준에서는 최소한 수작업색인만큼 좋다고 하였으며, 또한 편집된 MAI는 수작업색인과 비교해 재현율은 필적할만하고 정도율은 더 좋다고 하였다. 이 방법에 의한 색인법은 현재 NASA의 Center for Aerospace Information에서 사용되고 있다.[10]

2. 할당색인법(Assignment Indexing)

문헌으로부터의 단어와 또는 문구의 발췌는 컴퓨터가 비교적 잘 성취할 수 있는 작업이다. 자동발췌는 수작업발췌보다 한가지 분명한 이점이 있는데 그것은 철저하게 일관성이 있다는 점이다. 그러나, 대부분의 수작업색인은 발췌색인이 아니라 할당색인이며, 컴퓨터가 이 작업을 수행하는 것은 보다 어렵다. 컴퓨터로 할당색인을 작성하기 위한 명확한 방법은, 할당될 각 용어를 위해, 수작업색인자가 그 용어를 할당할 문헌에서 자주 출현할 경향이 있는 단어와 문구들의 "프로파일"을 개발하는 것이다. 예컨대, 'acid rain'을 위한 프로파일은 'acid rain', 'acid precipitation', 'air pollution', 'sulfur dioxides' 등과 같은 문구를 포함할 수 있다.

만일 통제어휘집에 있는 모든 용어가 그 용어와 연관된 프로파일을 갖는다면 컴퓨터 프로그램은 어떤 문헌내의 중요한 문구(근본적으로 앞에서 언급한 빈도수 기준에 의해 발췌된 문구)를 이 프로파일과 대조하여 문헌프로파일이 용어 프로파일과 어떤 기준치 이상으로 일치될 때 그 문헌에 특정 용어를 할당할 수 있을 것이다.

이 방법은 비교적 단순한 것 같지만 실행은 그렇게 쉽지 않다. 첫째, 일치의 기준(matching criteria)이 약간 복잡하다. 만일 'acid rain'이 논문에서 10번 출현한다면 이는 색인용어로서 거의 확실히 할당되어야 한다. 그러나 'acid rain'이 문헌에서 단지 두 번만 출현하고 'atmosphere', 'sulfur dioxide', 'sulfuric acid'가 모두 오히려 자주 출현하는 경우를 가정해 보면 과연 'acid rain'이 색인용어로 할당되어져야 하는가? 단어나 문구의 여러 가지 다른 조합이 특정한 색인용어가 할당의 후보가 되어야만 하는 사실의 동기가 될 수 있다는 것은 분명

8) Klingbiel, P. H, *Machine-aided Indexing*. Alexandria, VA. : Defense Documentation Center, 1971.
9) Klingbiel, P. H. and Rinker, C. C, "Evaluation of machine-aided indexing," *Information Processing & Management*, Vol. 12(1976), pp. 351-366.
10) Silvester, J. P. et al, "Machine-aided indexing at NASA," *Information Processing & Management*, Vol. 30(1994), pp. 631-645.

하다. 더욱이 특정용어가 할당되어야 하는 전조(前兆)로서 각 조합의 중요성은 상이한 동시출현값(co-occurence values)의 사용을 포함할 것이다. 예컨대, 만일 'beat', 'lake'와 'pollution'의 단어들이 모두 어떤 문헌에서 두세번 출현한다면 용어 'water pollution'은 물론 'thermal pollution'이 이 문헌에 할당될 수 있는 충분한 근거가 될 수 있다. 그러나 'beat'와 'lake'는 'pollution'의 출현없이는 두 용어가 한 문헌에 아주 많이 출현해야만 'thermal pollution'이 할당될 수 있는 좋은 후보용어가 될 수 있다. 또한 수작업색인자가 비교적 쉽게 할당할 수 있는 어떤 용어들이 컴퓨터에 의해서는 거의 할당되지 않는 경우도 있다.[11]

이와같은 문제들 때문에 용어를 자동적으로 할당하려는 초기의 시도들은 매우 적은 색인용어들이 어휘집에 포함된 경우 조차도 매우 성공적이지 못하였다.[12] 그러나 지난 30여년동안에 보다 훌륭한 컴퓨터 프로그램이 개발되었으며 현재는 보다 성공적인 할당색인이 가능하다.

Van der Meulen과 Janssen[13]은 자동할당색인과 수작업색인을 비교하였는데, 이들은 자동색인이 수작업색인만큼 좋은 결과를 달성하였다고 주장했다. 그러나 이 판정은 단지 두 개의 탐색결과에 근거한 것이다.

자동할당색인을 위한 보다 정교한 프로그램이 BIOSIS에서 개발되었다.[14] 논문의 표제에 나타난 단어들이 약 15,000개의 생물학 용어들로 구성된 의미사전(Semantic Vocabulary)에 대조되었으며, 이들은 차례로 600개의 개념표목(Concept Headings : 비교적 광범위한 주제표목) 사전에 결합되었다. 따라서 개념표목이 표제에 출현하는 단어/문구에 근거하여 컴퓨터에 의해 할당될 수 있었다. Vleduts-Stokolov는 수작업에 의해 할당된 개념표목들의 약 61%가 단지 표제에만 근거하여 컴퓨터가 할당할 수 있었다고 보고하였다. 만일 일차와 이차의 할당만을 고려한다면 (BIOSIS는 일차, 이차 및 삼차의 세 수준의 용어가중치 일람표를 사용하였다) 약 75%의 할당이 자동적으로 수행될 수 있었다. 그러나 실제는 프로그램이 그처럼 높은 성능수준을 달성하지 못하였다. 프로그램은 일차와 이차 할당에서 약 80-90%의 성공을 달성하였으며(표제에 근거하여 이론적으로 할당될 수 있었던 75%중에서 80-90%을 할당하였음), 삼차까지의 모든 할당에서도 거의 이 수준의 성공을 달성하였다(단지 표제에 근거하여 출현할 수 있었던 61%의 할당중에서 대략 80%내지 약간 더 좋은 수준을 할당하였음). 다른 단어들에서는 약간의 할당누락(underassignment) 이 발생하였다. 즉, 할당되어야하는 약간의 용어들을 프로그램은

11) O'Connor, J, "Automatic subject recognition in scientific papers:an empirical study," *Journal of the Association for Computing Machinery*, Vol. 12(1965), pp. 490-515.

12) Borko, H. and Bernick, M, "Automatic document classification," *Journal of the Association for Computing Machinery*, Vol. 10(1963), pp. 151-162.

13) Van der Meulen, W. A. and Janssen, P. J. F. C, "Automatic versus manual indexing," *Information Processing & Management*, Vol. 13(1977), pp. 13-21.

14) Vlenduts-Stokolov, N, "Concept recognition in an automatic text-processing system for the life sciences," *Journal of the American Society for Information Science*, Vol. 38(1987), pp. 269-287.

할당하지 못하였으나 수작업은 할당하였다. 동시에 약간의 할당초과(overassignment)도 또한 발생하였다. 즉, 할당돼서는 안되는 약간의 용어들이 할당되었다. 이것은 할당누락과 동일한 범위내에서 발생하였는데, 이는 컴퓨터에 의한 용어할당의 80-90%가 정확하였는데, 이는 수작업색인자들도 역시 동일한 수준의 용어할당을 하였다는 의미에서이다.

약간 유사한 방법이 1971-77년의 기간동안 ABI/INFORM(Business분야의 DB)에 있는 초록들을 자동으로 색인하기 위해서 사용되었다.[15] 거의 19,000 용어들로된 "중개어휘집" (bridge vocabulary)이 텍스트표현들을 통제어휘집의 용어들로 안내하기 위하여 개발되었다. 표제나 초록에 한번 출현하는 용어는 하나의 통제용어 할당에 충분한 요인이 되기 때문에 자동색인은 수작업색인보다 하나의 아이템에 보다 많은 용어를 할당하는 경향이 있었다.(아이템당 수작업색인의 8-12개에 비해 평균 16개)

American Petroleum Institute에서 수행된 MAI는 BIOSIS의 그것과 유사하다.[16] 목적은 초록에 근거하여 API시소러스의 통제용어를 컴퓨터가 할당할 수 있는 방법을 개발하는 것이었다. 시스템의 초기 버전은 수작업색인자가 할당한 용어들의 약 40%를 할당하였으며 또한 많은 원하지 않는 용어들을 할당하였다. 그러나 이 실험 이후 컴퓨터는 할당되어야 할 용어들의 약 80%를 할당할 수 있었으며 또한 원하지 않는 할당의 뚜렷한 감소를 가져왔다. 사실 최초의 실험이래 상당한 발전이 이루어 졌다. Martinez 등은[17] 시스템의 발전과 또한 텍스트표현들을 시소러스용어들로 결합하는데서 야기된 문제들을 기술하였으며, 또한 Hlava는[18] 한 언어의 색인용어를 다른 언어의 색인용어(예컨대, 영어를 독일어로, 독일어를 영어로)로 결합하는 API 방법의 개선을 검토하였다.

텍스트표현을 디스크립터로 결합(mapping)하기 위한 보다 정교한 방법이 Technische Hochschule Darmstadt에서 개발되었다.[19][20][21] Darmstadt 방법은 특수한 텍스트표현이 표제나 초록에서 출현하면 이 아이템에 디스크립터가 할당되는 확률을 산정하는 가중치 접근 방

15) Trubkin, L, "Auto-indexing of the 1971-77 ABI/INFORM database," *Database*, Vol. 2, No. 2(1979), pp. 56-61.

16) Brenner, E. H. et al, "American Petroleum Institute's machine-aided indexing and searching project," *Science and Technology* Libraries, Vol. 5, No. 1(1984), pp. 49-62.

17) Martinez, C. et al, "An expert system for machine-aided indexing," *Journal of Chemical Information Science*, Vol. 27(1987), pp. 158-162.

18) Hlava, M. M. K, "Machine-aided indexing(MAI) in a multilingual environment," *Online Information* 92, (1992), pp. 297-300.

19) Knorz, G, *Automatisches Indexieren als Erkennen abstrakter Objekte*. Tubingen : Max Niemeyer Verlag, 1983.

20) Fuhr, N, "Models for retrieval with probabilistic indexing," *Information Processing & Management*, Vol. 25(1989), pp. 55-72.

21) Biebricher, P. et al, "The automatic indexing system AIR/PHYS-from research to application," In : *Readings in Information Retrieval; ed by K. Sparck Jones and P. Willett*, pp. 513-517. San Francisco : Morgan Kaufmann, 1997.

법이다.

앞에서도 언급한 바와 같이 자동할당색인의 가장 성공적인 활용은 Klingbiel의 연구에 기초하여 NASA의 Center for Aerospace Information에서 현재 사용되고 있는 것이다.[22)]

자동할당색인이 지난 30여년 동안에 상당히 개선되었지만 대규모 어휘집(예컨대, 10,000개의 디스크립터가 포함된 시소러스)으로부터 인간의 개입(중재)없이 용어들이 완전자동으로 할당될 수 있는 수준에는 아직 도달하지 못하였다. Hersh 등의[23)] 연구는 의학분야에서 텍스트를 통제어휘집(Unified Medical Language System의 통제어휘집)의 용어들로 결합하여 얻은 결과보다 간단한 텍스트탐색의 결과가 보다 우수하다고 주장하였다.

실제로 자동할당색인은 현재 인쇄된 색인지의 생산을 제외하고는 별로 관심이 없는 상태다. 30년 전에 자동할당색인은 보다 많은 일반적 관심사였다. 그때는 대량의 텍스트를 컴퓨터로 축적하고 처리하는 것이 매우 많은 비용이 들었기 때문에 텍스트를 다소 단축하는 모든 방법이 정당화될 수 있었다. 이제는 물론, 만일 어떤 아이템의 전문(全文)이 전자적 형태로 존재하거나 또는 적당한 초록이 있다면 DB로부터 어떤 형태의 인쇄된 색인을 생산하지 않는다면 색인을 계획하는 것은 거의 의미가 없다. 그럼에도 불구하고 후술한 바와 같이 자동할당색인이 아직도 유용하다는 관점에서 특별한 활용법이 존재하고 있다.

특별한 형태의 인쇄색인으로는 도서의 권말색인이 있다. 컴퓨터로 이 유형의 색인을 생산하는 연구는 30여년 이전에 시작되었다. Artandi는[24)] 화학분야에서 컴퓨터로 도서색인을 생산하였으며, 이 방법으로 생산된 색인이 수작업색인과 비교하여 대등하였으나 많은 비용이 든다고 주장하였다. 그러나 대부분의 비용이 텍스트를 전자적 형태로 입력하는데 드는 비용이었다. 실제로 모든 인쇄가 이제는 전자적 입력으로부터 수행되기 때문에 비용요인들은 더 이상 큰 문제가 되지 않을 것이다. 그럼에도 불구하고 도서색인을 자동으로 생산하는 문제는 Artandi의 연구가 암시하는 것보다 훨씬 어렵다. 비록 한정된 주제분야라 할지라도 표현용어들의 대규모 어휘집이 필요할 것이며, 또한 각 용어에 대한 상당수의 가능한 결합문구도 많을 것이다. 더욱이 이들 두 어휘집(표현용어와 결합문구의 어휘집)이 해당분야에서의 새로운 발전과 변화되는 전문용어를 반영하기 위해 계속 갱신되어야 할 것이다.

물론 Artandi는 할당색인을 시도하였다. 텍스트로부터 색인기입으로 사용될 수 있는 적합문구의 발췌는 보다 용이한 문제다. Earl은[25)] 명사구의 발췌를 포함하는 도서색인의 생산방법을

---

22) Silvester, J. P. et al, 1994. op. cit.

23) Hersh, W. R. et al, "Words, concepts, or both : optimal indexing units for automated information retrieval," *Sixteenth Annual Symposium on Computer Applications in Medical Care,* (1993), pp. 644-648.

24) Artandi, S, *Book Indexing by Computer.* Doctoral thesis, New Brunswick, NJ, Rutgers : the State University, 1963.

25) Earl, L. L, "Experiments in automatic extracting and indexing," *Information Storage and Retrieval,* Vol. 6(1970), pp. 313-334.

기술하고 있다. 그녀는 만족스런 도서권말색인이 불필요한 용어들을 삭제하기 위한 사후 편집으로 자동으로 생산될 수 있었다고 주장한다. 뒤에 Salton은[26] 구문분석과정이 도서색인에 사용될 적합문구를 생성하기 위해 사용될 수 있는 방법을 기술하였다. 다른 한편으로 Korycinski와 Newell은[27] 도서색인의 자동생산이 논문의 자동색인보다 훨씬 어려운 이유들을 검토하였다.

대부분의 자동색인시스템은 컴퓨터가 인간을 대신한다는 의미에서는 진정한 자동이 아니지만 그러나 컴퓨터는 수작업색인자를 도우려고 한다는 의미에서 "machine-aided"라는 용어가 "automatic"이라는 용어보다 더 좋을 것이다.

일반적으로 MAI를 위한 두가지 중요한 접근방법은 다음과 같다.

1) 컴퓨터는 다양한 유형의 online display를 제공하기 위해 사용되며 색인자를 신속하게 도와준다. 색인자에 의한 오류(예컨대, 비표준 용어의 사용이나 쓸모없는 주표목/부표목의 조합)는 실시간에 인지되어 색인자에게 즉시 통고될 수 있다.
2) 컴퓨터프로그램은 텍스트(아마도 표제와/또는 초록만)를 읽고 발췌나 할당과정으로 색인용어들을 선정하는데 사용된다. 그러므로 선정된 용어들은 수작업색인자에 의해 체크될 수 있으며, 수작업색인자는 프로그램이 할당할 수 없었던 접근점들을 추가할 수 있고, 또한 잘못하여 할당한 용어들을 삭제할 수 있다.

Ⅲ. 인공지능과 전문가시스템

인터넷은 일반적으로는 정보검색기법에서 그리고 특별하게는 자동방법에서 엄청난 관심을 증가시켰다. 수년 전만해도 단지 실험적인 것으로 여겨졌던 약간의 시스템과 기법들이 이제는 상업적으로 활용되고 있다. 전문가시스템이나 인공지능기술을 색인작성 및 관련된 업무에 적용하기 위하여 근래에 많은 연구가 시도되고 있다. 먼저 전문가시스템 접근방법을 살펴보고, 다음으로 "지능적 텍스트처리"의 다양한 출현을 살펴보고자 한다.

색인업무가 보다 전문화 될 수록 그것은 더욱 진정한 전문가시스템 접근을 위한 비용대 효과면에서의 후보가 될 것이다. 즉, 텍스트나 또는 어떤 대용물을 효과적으로 색인하기 위하여

26) Salton, G. A, *Syntactic Approach to Automatic Book Indexing*. Ithaca, NY, Cornell University : Department of Computer Science, 1989.
27) Korycinski, C. and Newell, A. F, "Natural-language processing and automatic indexing," *The Indexer*, Vol. 17(1990), pp. 21-29.

그것들을 이해하고 범주화할 진정한 전문가들이 요구될 수 있다. 이러한 전문가들의 시간은 비쌀것이기 때문에 적어도 색인작성의 일부를 수행하는 면에서 낮은 수준의 인력(personnel)을 도와주는 시스템은 좋은 투자가 될 것이다. Swaby는[28] 미화석(microfossils)의 식별을 위해 설계된 시스템에 대해 비교적 상세하게 기술하였다. 전문가가 아닌 조작자(operator)가 대상물의 범주화 작업에 포함된다. 시스템은 식별작업을 돕기 위해 텍스트와 그래픽스의 조합을 사용한다는 점에서 흥미롭다. 이용자는 화석들의 상(image)과 그들의 결합된 속성(associated attributes)과 속성값을 함께 스크린에 보여주는 방법으로 화석을 식별한다. 이 시스템은 (의학적)진단시스템과 다소 유사하게 작용한다. 즉, 갖고 있는 화석에 적용하여 선정된 속성에 근거하여 시스템은 비전문가인 이용자에게 가장 적당한 식별을 하도록 한다.

Tway와 Riedel은[29] 어느 정도 유사한 시스템인 COREXPERT에 대해 기술하고 있는데, 이는 비전문가에게 해양 밑바닥으로부터 뽑아낸 침전물 채취샘플을 기술(근본적으로 "색인")할 수 있도록 설계된 것이다. 기술자인 이용자(technician users)는 이 작업을 돕기 위한 지식베이스(침전물 성분 및 다른 시각적 보조자료의 사진을 포함한)의 모든 특성을 사용해서 샘플의 다양한 특성을 기술하기 위한 데이터 기입 스크린(data entry screen)을 완성한다. 이 시스템은 "지능형 데이터 기입"시스템으로 불리어 지는데 어느 정도 수준의 지능이 오류를 피하기 위해 수립된다. 예를 들면, 시스템은 입력된 데이터에서 변칙적인 것을 -예컨대, 특정 광물에 대해 변칙적으로 높거나 낮은 선광(選鑛)이나 또는 정상적으로는 동시발생을 하지 않는 두 광물의 동시출현 등- 체크한다. 더욱이, 기술자는 샘플에서 발생하는 모든 광물 유형이나 화석집단을 직접 식별할 필요는 없고 다만 가장 중요한 것만 식별한다. 프로그램은 나타날 것으로 알고 있는 주요요소나 또는 수심(水深)이나 지리적 위치 등의 기록된 상황에 근거하여 다른 요소의 발생확률을 이끌어낼 수 있다. COREXPERT는 Scripps Institution of Oceanography를 위해 개발되었지만 사용이 중단되었다. 그 이유는 부분적으로는 자금부족과 또한 시스템을 사용하는 기술자들이 이 프로그램없이도 운영할 수 있는 충분한 지식이 있었기 때문이다. 이와 같은 사용중단은 활용 중에 있는 전문가시스템에서 꽤 자주 발생한다.

도서, 논문 및 기타 출판물에 적용된 온라인 MAI에 관한 연구는 30여년 전으로 거슬러 올라간다.(예컨대, Bennett[30] 및 Bennett 등의[31] 연구 등) 온라인 보조는 여러 가지 형태를 취할

---

28) Swaby, P. A, "Integrating artificial intelligence and graphics in a tool for microfossil identification for use in the petroleum industry," In : *Innovative Applications of Artificial Intelligence2* ; ed. by A. Rappaport and R. Smith, pp. 203-218. Cambridge, MA : MIT Press, 1991.

29) Tway, L. E. and Riedel, W. R, "Intelligent data entry," *PC AI*, Vol. 10, No. 1(1996), pp. 16-21.

30) Bennett, J. L, "On-line access to information : NSF as an aid to the indexer/cataloger," *American Documentation*, Vol. 20(1969), pp. 213-220.

31) Bennett, J. L, et al. *Observing and Evaluating an Interactive Process : a Pilot Experiment in Indexing*. San Jose, CA : IBM Research Laboratory, 1972.

수 있다. 예를 들면, 컴퓨터가 처리한 표제, 초록 또는 텍스트에 근거하거나 또는 색인자가 미리 입력한 용어에 근거하여 색인자에게 용어를 제시하는 형태, 시스템 어휘집에 없는 용어나 쓸데없는 용어조합 등의 확실한 색인자의 오류에 대한 신호형태, 허용되지 않은 용어를 허용된 용어로 대체해 주는 형태 및 어떤 용어들이 과거에 사용되었으며 또는 어떤 아이템들이 과거에 색인되었는가를 색인자가 찾아볼 수 있도록 해주는 DB와의 인터페이싱 형태 등이다.

현재 운영중인 온라인 색인시스템은 다양한 수준의 도움과 정교함을 제공한다. 예를 들면, NLM에서 현재 사용중인 AIMS(Automatic Indexing and Management System)는 다음과 같은 성능을 포함하고 있다 : 색인자가 입력한 모든 표목에 대해서, 명령을 받으면, 어떠한 부표목이 사용될 수 있는가를 보여주고, 용어주석과 범위주기를 보여주고, 바람직하지 못한 용어나 약어를 바람직한 것으로 전환하기 위한 상호참조를 사용하고, 통제어휘의 일부를 계층적으로 또는 자모순으로 보여주고, 색인자가 선택할 수 있는 "check tags"(일반적으로 적용할 수 있는 용어들의 소규모 집합)를 보여준다.

보다 정교한 MAI시스템은 어떤 아이템을 부분적으로 색인하거나 또는 적어도 색인자에게 용어를 제시하는 점에서 보다 우수하다. 하나의 예로서 CAIN이 진행중인 농업연구프로젝트의 EC(유럽공동체)DB인 AGREP를 사용하기 위해서 개발되었다. 프로젝트 기술사항은 표제, 초록 그리고 프로젝트 범위를 지시하는 비 통제용어들을 포함한다. CAIN은 이 텍스트를 두 개의 통제어휘집(AGROVOC와 CAB시소러스)에 대조하여 이들 어휘집으로부터 후보 용어들을 제시한다.[32]

짧은 텍스트(예컨대, 케이블)와/ 또는 비교적 적은 통제어휘집을 사용하는 시스템은 수작업 색인자가 필요한 정정이나 추가를 만들기 위해 검토하기 전에 많은 색인작성을 정확하게 수행할 수 있다.

대규모로 충분하게 운영되고 있는 MAI시스템은 NASA의 CAI(Center for Aerospace Information) 시스템이다.[33] 항공우주과학문헌에서 흔히 출현하는 지식기반 문구들이 NASA 시소러스의 용어들로 결합되도록 사용된다. 즉, 입력된 텍스트에서 이러한 문구들이 출현하면 시스템은 색인자가 검토(reviews) 할 수 있도록 후보디스크립터 리스트를 생산한다. 다른 기관에서 다른 어휘집을 사용하여 레코드에 할당한 용어들을 NASA 시소러스에 있는 용어들로 결합(mapping)하기 위한 방법이 CAI에서 연구 개발되었다.[34]

특별히 생의학분야에서 소규모의 전문화된 활용을 위한 상당히 흥미로운 자동색인을 찾아

---

32) Friis, T, "Assisted Indexing(CAIN)," *IAALD Quarterly Bulletin*, Vol. 37(1992), pp. 35-37.

33) Silvester, J. P. et al. "Machine-aided indexing at NASA," *Information Processing & Management*, Vol. 30(1994). pp. 631-645.

34) Silvester, J. P. et al. *Machine Aided Indexing from Natural Language Text. Status Report.* Linthicum Heights, MD, RMS Associates, 1993.

볼 수 있는데, 환자의 배설물 개요(discharge summaries)의 텍스트가 적합한 임상 디스크립터를 자동으로 할당하기 위하여 분석된다.[35)]

특수한 주제분야에서 이러한 유형의 할당색인이 적절하게 수행되고 있지만, 자동처리방법들이 일반적으로 수작업색인자가 달성한 수준에는 미치지 못하고 있다.[36)] 그럼에도 불구하고 이러한 유형의 자동색인은 예비할당을 수행함으로써 수작업색인자의 수고를 경감시켜 줄수 있다.

Rindflesch와 Aronson은[37)] 텍스트를 의학용어집(Unified Medical Language System의 의학용어집)으로 결합하는데 내포된 약간의 애매모호한 문제들을 검토하고 이를 제거하는 몇가지 규칙을 제시하였다.

Lirov와 Lirov는[38)] 특수한 주제서지의 생산을 도와주는 전문가시스템을 기술하였다. 일단 온라인 DB로부터 서지레코드가 검색되어 다운로드되면 시스템은 저자와 주제색인의 생산을 도와준다 : 즉, 키워드들이 서지레코드에 있는 디스크립터들에 대조되어 키워드에 의한 서지가 배열된다.

복잡한 주제분야(예컨대, 의학, 화학 또는 물리학 등)를 취급하고 있는 논문길이 정도의 텍스트를 위한 완전자동할당색인(인간의 개입이 없는)은 실현이 요원하다. 특히 사용되는 통제어휘집이 아주 대규모 일 때는 색인자를 돕기 위한 보다 정교한 전문가시스템의 생성을 위한 연구가 계속되는데 괄목할만한 예로서 NLM이 수년동안 개발하고 있는 MedIndex를 들 수 있다.[39)] 이것은 전형적인 프레임기반 전문가시스템이다. 경험있는 색인자일 필요는 없지만 적어도 의학문헌과 의학전문용어에 대한 약간의 이해가 요구되는 사용자는 여러 가지 적합한 프레임(예컨대, 질병의 유형, 치료법의 유형 등)으로 안내되어 그들을 효과적으로 완성하도록 촉구된다. 시스템은 색인자에게 특정 용어를 할당하도록 촉구할 수 있으며, 또한 용어가 부적

---

35) Borst, F. et. al, "TEXTINFO : a tool for automatic determination of patient clinical profiles using text analysis," *Fifteenth annual Symposium on Computer Applications in Medical Care,* pp. 63-67. New York : McGraw Hill, 1992.

36) Chute, C. G. and Yang, Y, "An evaluation of concept based latent semantic indexing for clinical information retrieval," In : *Sixteenth Annual Symposium on Computer Applications in Medical Care,* pp. 639-643. New York : McGraw Hill, 1993.

37) Rindflesch, T. C. and Aronson, A. R, "Ambiguity resolution while mapping free text to the UMLS metathesaurus," In : *Eighteenth annual Symposium on Computer Applications in Medical Care,* pp. 240-244. Philadelphia : Hanley & Belfus, 1944.

38) Lirov, Y. and Lirov, V, "Online search + logic programming = subject bibliography : an expert systems approach to bibliographic processing," *Online Review,* Vol. 14, No. 1(1990), pp. 3-12.

39) Humphrey, S. M, "Interactive knowledge-based systems for improved subject analysis and retrieval," In : Artificial Intelligence and Expert System : *Will They Change the Library? ed by F. W. Lancaster and L. C. Smith,* pp. 81-117. Urbana-Champaign : University of Illinois, Graduate School of Library and Information Science, 1992.

당하게 사용된 때에는 색인자에게 정정하도록 할 수 있다. 예를 들면, 질병의 위치(예컨대, 뼈 종양)를 반영하는 종양(암) 용어를 할당한 색인자는 종양의 조직학적 유형(예컨대, 선종암)을 나타내는 동반용어(companion term)를 할당하도록 깨닫게 될 수 있다. 또는 "femur and bone neoplasms"(대퇴골과 뼈종양)과 같은 부적당한 조합을 할당한 색인자는 정확한 용어 "femoral neoplasms"(대퇴부의 종양)를 통보받을 수 있다. MedIndex는 여전히 시험중이므로 아직까지는 충분히 평가되지 못하였다.

매일매일의 색인작성을 돕는 것보다는 오히려 색인자들의 훈련을 돕기위한 전문가시스템들이 개발되고 있는데, 이 유형의 하나가 NAL(미국 국립농업도서관)에 있는 CAIT(Computer-Assisted Indexing Tutor)이다.[40]

주제색인작업을 도와주는 모든 컴퓨터시스템은 전문가시스템(적어도 전문가시스템이란 용어의 가장 허술한 의미에서)으로 간주될 수 있으며, 만일 컴퓨터시스템이 경험이 부족한 사람을 전문색인자의 작업에 접근할 수 있도록 도와준다면 이 의미는 특히 강조될 수 있다. 그리고 색인자에게 용어를 시사하거나 또는 색인자의 오류를 정정하는 시스템들은 적어도 "지능"의 소량(modicum)을 제공하는 것으로 간주될 수 있다.

문헌에 기술된 약간의 시스템이나 프로그램들은 "인공적으로 지적"인 것으로 간주된다. Driscoll 등과[41] Jones와 Bell이[42] 좋은 사례를 기술하고 있는데, Jones와 Bell이 기술한 시스템은 색인기입을 작성하기 위하여 텍스트로부터 단어나 문구들을 발췌한다. 이 시스템은 주로 컴퓨터에 축적된 리스트에 근거하여 작업한다. 리스트는 두 종류가 있는데 하나는 단수/복수 형태를 결합하기 위하여, 그리고 간단한 구문분석(parsing)을 허용하기 위하여 무시되어야 할 단어들과 알고 있는 중요한 단어/문구/이름 들의 리스트이며, 다른 하나는 동형이의어를 명확하게 구분해 주는 리스트인데 이들은 사전을 작성하기 위하여 결합된다.

Driscoll 등이 기술한 시스템도 텍스트에서 유용한 색인용어를 찾는 것이다. 텍스트는 3,000개 이상의 문구로 구성된 리스트와 대조된다. 텍스트에서 이들 중 하나가 출현한다면 삽입과 삭제 규칙의 사용을 개시한다. 삭제규칙은 단지 애매모호한 단어나 문구의 계속처리를 회피하는 것이고, 삽입규칙은 암시에 의해 찾아진 용어들의 한정된 집합("템플릿"을 완성하기 위하여)을 생성할 수 있다. 예를 들면, 'time', 'over'와 'target'란 단어들은 만일 그들이 서로 X개의 단어 이내에서 나타나면 'AIR WARFARE'를 생성할 것이다. Malone 등은[43] 이 시스템의

---

40) Irving, H. B, "Computer-assisted indexing training and electronic text conversion at NAL," *Knowledge Organization*, Vol. 24(1997), pp. 4-7.

41) Driscoll, J. R. et al, "The operation and performance of an artificially intelligent keywording system," *Information Processing & Management*, Vol. 27(1991), pp. 43-54.

42) Jones, K. P. and Bell, C. L. M, "Artificial intelligence program for indexing automatically(AIPIA)," *Online Information 92*, (1992), pp. 187-196.

43) Malone, L. C. et al, "Modeling the performance of an automated keywording system," *Information*

성능을 예측하기 위한 통계적 모델을 제시하였다.

Driscoll등과 Jones와 Bell이 기술한 유형의 시스템들은 정교하다. 그들은 수작업색인자들이 달성한 수준과 필적하는 발췌색인이나 또는 제한된 할당과 더불어 발췌를 수행할 수 있는데 비용은 더 저렴하다. 최소한 그들은 사람의 검토를 위한 후보용어들을 생산하는데 유용하다.

1997년 Dow Jones사는 다우존스 뉴스/검색서비스내에 있는 단체명의 자동인식과 표준화를 위한 "지식색인시스템"을 소개하였다. 1997년 5월 현재 약 40,000개의 단체명이 식별될 수 있고, 이용자는 이 이름들의 모든 정당한 변형을 탐색하는 능력을 갖고 있다. 이 시스템은 매일 37,000개의 기사까지 색인할 수 있다고 주장한다.

Woodruff와 Plaunt는[44] 자동지리색인을 위한 참신한 시스템을 기술하였는데, 텍스트에서 식별된 지명이 그의 위도/경도 좌표와 또한 삼림, 저수지, 항구 및 늪과 같은 연관된 특징을 보충하는 DB에 대조될 수 있다.

Ⅳ. 인터넷과 색인작성

인터넷을 통해 이용할 수 있는 엄청난 양의 정보자원이 색인작성 및 색인작성보다 범위는 작지만 초록작성의 문제를 제기하고 있음은 분명한 사실이다. 따라서 인터넷의 정보자원을 색인하기 위해 현재 이용되고 있는 방법을 살펴보고자 한다.

1. 탐색엔진(Search Engines)[45]

탐색엔진은 인터넷 자원에 대한 색인을 구축하여 운영된다. 사실상 이는 텍스트로부터 단어나 혹은 문구들을 발췌하여 부울방법(때로는 다른 방법)을 사용하여 이들 발췌들의 효율적 탐색을 허용하는 파일들을 구축하는 것을 의미한다. 원칙적으로 이러한 탐색가능한 파일들은 1960년대 초에 레코드의 순차적 탐색을 대신하기 시작한 직접탐색 이래로 정보검색을 촉진하기 위해 사용된 전형적인 도치파일들에 불과한 것이다. 그러나 현대의 컴퓨터 능력은 30여년

Processing & Management, Vol. 27(1991), pp. 145-151.

44) Woodruff, A. G. and Plaunt, C, "GIPSY : automated Geographic indexing of text documents," *Journal of the American Society for Information Science*, Vol. 45(1994), pp. 645-655.

45) 탐색엔진들에 대한 보다 최신의 정보는 Search Engine Watch로부터 얻을수 있는데, 이의 웹사이트는 http://searchenginewatch.com이다.

전의 그것들과 비교해서 엄청나게 규모가 큰 도치파일들의 효율적인 갱신과 탐색을 허용하고 있으며 탐색방법도 또한 보다 세련되고 있다.

탐색엔진이 사용하는 도치색인은 자동으로 구축된다. 소프트웨어 장치들이 기존의 정보자원에 대한 갱신은 물론 색인할 자원을 찾아 웹을 돌아다닌다. 적합한 정보자원을 발견하면 색인하며, 선정된 정보자원에 관련있는 다른 정보자원들도 또한 파악하여 색인한다. 불행히도 이를 의미하는 전문용어가 불명확하다. 이러한 순회로보트(roving robots)들 (보다 정확하게는 'softbots')이 "crawlers", "spiders", "wanderers" 및 "worms" 등의 다양한 이름으로 불리워지고 있다.

탐색엔진들은 정보자원에서 무엇을 발췌하여 색인에 포함하며 또한 이러한 색인을 탐색하기 위해 어떤 능력을 제공하는가? 라는 점에 있어 매우 다양하다.

약간의 탐색엔진 생산자들은 정보자원에 대해 실제로 얼마나 많이 색인하고 있는가 (예컨대, X개의 문자까지 또는 최초의 Y개의 단어까지 등)에 대해 꽤 명백한 반면에 다른 것들은 이점에서 상당히 애매모호하다. 어떤 경우에는 탐색엔진이 웹페이지 구축자에게 키워드와/ 또는 설명적 텍스트(descriptive text)를 포함하는 "metatag" 필드를 구축하여 무엇을 색인할 것인지를 결정하는 능력을 제공하고 있다. 대부분의 탐색엔진은 도치파일을 구축하기전에 주제내용을 나타내지 못하는 단어들을 제거하기 위해 전통적인 불용어리스트를 적용하고 있다.

상이한 탐색엔진이 상이한 탐색능력을 제공하지만 비교적 공통되는 탐색능력을 살펴보면 다음과 같다.

- 전형적인 부울논리의 사용
- 정확한 문구대조(phrase match)
- 특정집합 중에서 부분집합을 지정하는 능력(예컨대, 집합 "개"에서 "복서"만 찾아내는 능력)
- 레코드의 특정부분에 탐색을 한정하는 능력(예컨대, 표제나 또는 URL의 구성요소)
- 단어의 절단기법
- 단어 인접탐색
- 탐색예시에 의한 질의 (예컨대, 이미 유용할 것이라고 알려진 정보원과 유사한 정보원을 파악하는 능력)

어떤 탐색엔진은 보다 진보된 탐색능력을 주장한다. 예를 들면, Iconovex사의 생산제품에 적용되고 있는 Syntactica엔진은 텍스트의 자동색인작성에서 상당히 정교한 수준의 언어적 처리과정(예컨대, 텍스트내의 중요개념을 결정하기 위해 구문규칙과 어의적 가중치를 사용함)을 사용하고 있다.[46] 또한 Excite는 "실제의 키워드는 물론 개념들의 상호의존관계(corelation)를 토대로 하여 문헌들을 찾아 점수를 주기 위하여" 지능적 개념발췌(Intelligent Concept

46) http://www.iconovex.com/SWAPI/SWAPIWP.HTM

Extraction)를 사용한다고 주장하고 있다. 불행하게도 Excite사가 "우리의 신안특허출원중의 기술에 대한 상세한 운영은 기밀이라고 언급하고 있어 알수는 없지만 사용된 방법은 벡터에 근거한 용어의 가중치부여와 SMART시스템에서 Salton에 의해 처음으로 사용된 아이템의 그룹화 방법과 약간의 유사성이 있다고 보여진다. 그럼에도 불구하고 Venditto에[47] 의한 평가는 이러한 개념탐색의 사용과 키워드 탐색의 사용간에 거의 차이점이 없다고 하였다.

대부분의 탐색엔진이 탐색결과에 순위를 부여하고 있다. 물론 순위부여의 능력은 사용된 색인작성절차와 탐색파일에 축적된 정보의 양과 유형에 의존한다.

하나의 순위부여과정은 탐색식에서 사용된 단어들의 텍스트내 출현빈도수에 근거하여 선정된 정보원에 점수를 주는 것이다. 다른 대안은 일치되는 탐색어 수에 의해 선정된 아이템들을 순위화하는 방법이다.

어떤 경우에는 순위화가 몇가지 기준에 근거하고 있다. 예를 들면, Altavista는 다음과 같은 기준에 근거하여 점수화하는 기법을 사용한다고 주장한다.

- 탐색어가 아이템의 어디에 나타나는가?(표제나 각 데이터의 머리표제(header)에서와 같이 아이템의 앞부분에 나타나는 단어들은 보다 높은 가중치를 얻는다)
- 탐색어들이 서로 얼마나 가까운가?
- 정보원에서의 탐색어/탐색문구들의 출현빈도수

그러나 이러한 다양한 요소들이 점수화에 있어 어떻게 조합되는지 또는 어떤 요소에 가장 큰 가중치가 주어지는지에 대한 지시는 없다.

어떤 탐색엔진은 다른 것보다 좋은 출력능력을 보여준다. 예를 들면, Excite는 "자동적 주제 그룹화"를 제공한다. Excite에서는 검색된 아이템들이 문맥(context)에 의해 자동으로 그룹화된다. 따라서 단어 'bond'에 의해 검색된 텍스트들은 화학, 기계학, 심리학, 재정학 및 기타의 그룹으로 분리되어 질 수 있다.

인터넷에서의 탐색엔진의 확산은 수 개의 고급탐색엔진(metasearch engines)의 출현을 유도하고 있다. 고급탐색엔진은 탐색식을 한번에 수 개의 엔진에 제공한다. 고급탐색엔진은 수 개의 탐색엔진을 위해 결과를 조합할 수 있고 그리고 어떤 엔진이 어떤 아이템을 검색하였는지를 보여준다. 이러한 도구들의 가장 분명한 가치는 매우 완벽한 결과를 요구하는 탐색 수행에 있다. 그러나 일반적으로 검색된 아이템 수를 제한한다는 사실 때문에 이것은 부정적이다. 그들은 탐색결과를 이용자에게 제공하는데 있어서 개별 엔진보다 분명히 늦다. 아마도 그들의 가장 유용한 활용은 포함된 엔진들의 비교와 평가를 허락하는 하나의 도구를 제공한다는 점에 있다. 가장 잘 알려진 고급탐색엔진의 하나인 Savvy Search는 19개의 탐색엔진을 갖고 있는데 탐색식과 다양한 다른 요소들에 근거하여 이 엔진들을 순위화 한다고 주장한다.

---

47) Venditto, G, "Search engine showdown," *Internet World,* Vol. 7, No. 5(1996), pp. 79-86.

Jakob은[48] 탐색엔진들의 평가를 위한 기준을 제시하고 있는데, 그에 의하면 이상적인 엔진은 :

- 사용하기 쉬워야 하고(그러나 질문형성을 위한 도움을 포함해야 함)
- 신속한 방법으로 탐색을 수행해야 하고
- 기본적인 탐색과 더불어 보다 복잡한 부울탐색을 지원해야 하고
- 부분적인 단어탐색을 수행하고, wildcards(임의문자기호)의 사용으로 절단기법을 허용해야 하고
- 문구탐색과 인접탐색을 허용해야 하고
- 다양한 탐색용어에 대한 가중치 적용을 이용자에게 허용해야 하고
- 대 · 소문자 구별(case-sensitive)탐색에 대한 통제를 이용자에게 허용해야 하고
- 가장 좋은 탐색용어를 결정하기 위한 시소러스가 구축되어야 하고
- 최대한의 검색건수 선택을 이용자에게 허용해야 하고
- 제한된 수의 요소를 색인하는 것보다 오히려 문헌의 전문 색인작성을 해야하고 (비록 이용자가 어느 필드들이 탐색을 제한하기 위해 사용되는 가를 통제해야만 하더라도)
- 각각의 검색(hit)에 대한 표제와 URL을 포함하는 정보적 결과집합(informative result set)을 제시해야 하고
- 쉽게 설명된 결과집합을 적합성 점수나 또는 검색에 대한 순위부여시스템과 함께 제시해야 하고
- 정보원이 언제 색인되었는가를 제시해야 하고
- 시효가 지난 불필요한 링크(outdated null links)를 제거하기 위해 정규적으로 DB를 갱신해야 하고
- DB에 포함되지 않은 URL의 등록을 이용자에게 허용해야 한다.

현존하는 어떠한 탐색엔진도 이러한 모든 요구사항을 만족시켜 주지 못하며 또한 어느 하나의 탐색엔진이 할 수 있을 것 같지도 않다.

탐색엔진들에 대한 수차례의 평가나 비교가 수행되어 보고되었는데, 무엇보다 중요한 것은 매우 특수한 탐색에서 조차도 엄청나게 많은 아이템이 검색될 수 있음을 강조하고 있는 점이다. 예를 들면, Crohn 질병에 관한 정보를 검색한 Falk는[49] 'Crohn'이라고 하는 단 하나의 키워드가, 사용된 탐색엔진에 따라, 500에서 6,000 사이의 아이템을 검색하였다. 'Crohn'을 'drug' 또는 'surgery'와 조합해도 규모가 큰 엔진에서는 여전히 1,000 또는 그 이상의 아이템을 검색하였다. Altavista에서는 Crohn이란 한 단어에 대한 탐색이 가장 최근의 한달 동안의

48) Jakob, D, *Finding information on the Web.* October 10, 1995. http://www.nlc-bnc.ca/publications/netnotes/notes15.htm

49) Falk, H, "World Wide Web search and retrieval," *Electronic Library,* Vol. 15(1997), pp. 49-55.

입력(정보)에 한정하였는데도 800개 아이템을 검색하였다.

Scales와 Felt,[50] Conte,[51] Kimmel,[52] Venditto,[53] Pfaffenberger,[54] Chu와 Rosenthal,[55] Ding과 Marchionini,[56] Watson 등,[57] Dong과 Su,[58] 및 Nicholson도[59] 탐색엔진의 능력을 산출하였는데, 이 산출값들이 너무 빨리 변화되어 이들이 출판되었을 때에는 이미 약간 뒤떨어진(쓸모없는) 계수가 되어 버린다.

30여년 이상 정보검색분야에 종사해온 사람들을 가장 놀라게 한 것은 매우 특수한 탐색에서 조차 상례적으로 수천 아이템을 검색하는 도구들에 대한 광신적 표현이다. 아마도 이들 대부분은 매우 저질이거나 또는 여분(다른 아이템이나 정보의 복제)의 것으로 완전히 부적합할 것이다.

탐색절차가 더욱 "자동적"으로 보여질수록 이용자들은 보다 덜 비판적인 것 같다. 정보검색도구로서의 인터넷에 대한 현재의 광신은 대다수의 이러한 시스템들이 오히려 불충분하게 수행한다는 사실에도 불구하고 오랜 기간동안 여전히 고집하고 있는 의학분야의 기계보조 진단에 대한 광신을 약간 닮고 있다. 예를 들면, 4개의 이러한 시스템을 평가한 Berner 등은[60] 정확한 진단이 순위화된 출력에서 좀처럼 상위에 나타나지 않고 비교적 하위수준에 자주 나타나는 것을 발견하였는데, 이는 이러한 시스템들은 경험이 없는 의사에게는 실제로 위험스러울 수 있음을 암시하는 것이다.

Srinavasan 등은[61] 웹내에서의 색인작성과 아이템의 검색에 연관된 문제들을 연구하였는데, 이들의 연구는 보다 안정적인 환경(예컨대, 검색된 아이템의 순위화를 위한 역 용어빈도수)에

---

50) Scales, B. J. and Felt, E. C, "Diversity on the World Wide Web : using robots to search the Web," *Library software Review,* Vol. 14(1995), pp. 132-136.

51) Conte, R., Jr, "Guiding lights," *Internet World,* Vol. 7, No. 5(1996), pp. 40-44.

52) Kimmel, S, "Robot-generated databases on the World Wide Web," *Database,* Vol. 19, No. 1(1996), pp. 40-43, pp. 46-49.

53) Venditto, G, op. cit.

54) Pfaffenberger, B, *Web Search Strategies.* New York : MIS Press, 1996.

55) Chu, H. and Rosenthal, M, "Search engines for the World Wide Web : a comparative study and evaluation methodology," *Proceedings of the American Society for Information Science,* Vol. 33(1996), pp. 127-135.

56) Ding, W. and Marchionini, G, "A comparative study of Web search service performance," *Proceedings of the American Society for Information Science,* Vol. 33(1996), pp. 136-142.

57) Watson, J. et al, "Internet text retrieval : benchmark study," *Inform,* Vol. 10, No. 4(1996), pp. 24-45.

58) Dong, X. and Su, L. T, "Search engines on the World Wide Web and information retrieval from the Internet : a review and evaluation," *Online & CDROM Review,* Vol. 21, No. 2(1997), pp. 67-81.

59) Nicholson, S, "Indexing and abstracting on the World Wide Web : an examination of six web databases," *Information Technology and Libraries,* Vol. 16(1997), pp. 73-81.

60) Berner, E. S. et al, "Performance of four computer-based diagnostic systems," *New England Journal of Medicine,* Vol. 330(1994), pp. 1792-1796.

61) Srinavsan, P. et al, "An investigation of indexing on the WWW," *Proceedings of the American Society for Information Science,* Vol. 33(1996), pp. 79-83.

서 만족스럽게 일할 수 있는 기법이 "이질적이고 동적인 문맥"(heterogeneous and dynamic context)에서는 덜 효과적일 수 있다고 시사하였다.

2. 정보 에이전트(Information Agents)

탐색엔진외에도 몇가지 유형의 "지적 에이전트"들이 인터넷상에서 다양한 기능을 수행하기 위하여 개발되었다. 탐색엔진 색인을 구축하기 위해 사용되는 crawler들이 이러한 도구들의 예시가 된다.

Blake는[62] "정보에이전트"를 "인간의 어떠한 개입 없이 다양한 정보자원으로부터 데이터를 재치있게 검색할 수 있는 소프트웨어"로서 정의하였다. Roesler와 Hawkins는[63] 이러한 에이전트들의 특성에 관해 유용한 검토를 하였다.

여기에서는 주로 인터넷 사용을 위해 개발된 에이전트만을 언급하고자 한다. 현재 소프트웨어도 상업적으로 또는 적어도 실험적 토대 위에서 이용가능한데 이들은 인터넷 자원에 관련된 다음과 같은 다양한 업무를 수행하기 위해서 개발되었다.[64]

- 가장 저렴한 가격의 생산물(예컨대, CD)을 찾는 일
- 미리 설정된 이용자 기준에 따라 e-mail 메시지와 기타 아이템들을 여과(filtering)하는 일
- 불필요한 것이 될 수 있는 원치 않은 인터넷 광고와 기타 아이템들(어린이들의 접근이 부적당하다고 판단될 수 있는 것들 포함)을 여과하여 삭제하는 일
- 선정된 네트워크 자원을 감시하여(monitoring) 특수조건이 충족될때나 또는 단순히 자원이 갱신되었을 때 이용자에게 알려주는 일; 어떤 경우에는 에이전트가 변경이나 새로운 정보를 검색하여 이용자의 주의를 불러일으켜 준다.
- 이용자가 사용한 최초의 용어들에 동의어적이거나 거의 동의어적인 용어들을 찾아줌으로써 탐색질의를 확대하는 일
- 선정된 범주(예컨대, 경기 스코어, 금융 뉴스 등)내에서 네트워크 이용자에게 뉴스아이템이나 기타 유형의 정보를 "후원하는"(pushing)일
- 다양한 인터넷 자원이나 또는 기업의 인트라넷 범위내의 이종(異種)의 자원을 탐색하는 일
- 관심 프로파일에 근거하여 이용자의 마음에 드는 형태(필름, CD, 도서 등)를 제공하는 일

---

62) Blake, P, "Information agents," *Online & CDROM Review*, Vol. 18(1994), pp. 189-190.
63) Roesler, M. and Hawkins, D. T, "Intelligent agents," *Online*, vol. 18, No. 4(1994), pp. 19-32.
64) Lancaster, F. W, *Indexing and Abstracting in Theory and Practice*. 2nd ed. Champaign, Ill : University of Illinoise, 1998. pp. 306-307.

- 인터넷 상에서의 "데이터 채집"(data mining)
- 웹서버 상에서 유지되는 지역 DB에 대한 색인작성

3. 문제점

웹기반 에이전트들은 필연적으로 문제점이 있다. 예를 들면 이들은 이미 과부하(overloaded) 상태일 수 있는 서버들에게 실질적인 별도의 부담을 초래할 수 있다. 또한 어떤 사이트들은 에이전트에 의한 접근을 이미 봉쇄하기 시작하였다. Eichmann은[65] 이러한 문제들을 검토하고 웹내에서의 에이전트 사용을 위한 "윤리"를 제안하였다.

잠재적으로 심각한 문제는 인터넷 내에서의 품질관리의 결여와 웹사이트의 개발자가 그들의 사이트가 발견되어지기를 원하는 사실과 연관되어 있는데, 이는 특히 이윤추구의 관심에서는 더욱 심하다.

그러나 가장 큰 인터넷 문제는 그의 절대적인 크기와 중요한 품질 여과작업의 결여이다. Kelly와 Nicholas는[66] 인터넷 이용과 관련된 몇가지 주요 문제점들을 확인하였는데, 이들은 너무 과다한 정보; 부족한 질, 적합성 또는 신빙성의 정보; 불충분한 조직 등이다. 불행하게도 상황은 개선되기보다는 오히려 더욱 악화되고 있는 것 같다. 국가정보하부구조(NII)을 취급하고 있는 미국인공지능협회는 이 상황을 다음과 같이 보고하고 있다 : "반도체 기억밀도, 처리속도 및 네트워크 대역너비에 있어서의 현재의 추세는 하부구조가 인터넷과 같은 현존 시스템보다 수천배나 크게 될 것을 암시하고, 따라서 국가정보하부구조가 지원해야 할 서비스도 상상할 수 없을 만큼 방대해 질 것이다."

Fairthorne은[67] "색인작성은 정보검색의 기본 문제이며 가장 비용이 많이 드는 장애"라고 컴퓨터가 이 문제에 어느정도 적용되기 이전에 이미 지적하였다. 인터넷자체의 규모가 Fairthorne의 논평이 있었던 40여년 전에 예견할 수 있었던 어느 것보다도 훨씬 큰 규모의 색인작성의 과제를 제시하고 있다.

Wellisch는[68] "전자잡지는 텍스트의 불안정 때문에 색인될 것 같지 않다"고 주장하였다. 인터넷상의 대부분의 자원이 잡지보다 훨씬 안정적이지 못하기 때문에 그는 아마도 자주 변화

---

65) Eichmann, D, "Ethical web agents," *Computer Networks and ISDN System*, Vol. 28(1995), pp. 127-136.
66) Kelly, S. and Nicholas, D, "Is the business cybrarian a reality? Internet use in business libraries," *ASLIB Proceedings*, Vol. 48(1996), pp. 136-144.
67) Fairthorne, R. A, "Automatic retrieval of recorded information," *Computer Journal*, Vol. 1, No. 1(1958), pp. 36-41.
68) Wellisch, H. H, "Book and periodical indexing," *Journal of the American Society for Information Science*, Vol. 45(1994), pp. 620-627.

되는 텍스트의 색인작성과 같은 모든 계획이 취소된 이유일 것이라고 감지하였을 것이다. Weinberg은[69] 자원의 규모와 불안전성의 이유로 문제의 복잡성을 전적으로 과소평가하면서 도서의 권말색인구조에 입각한 수작업색인작성이 해결책이 될 수 있을 것이라고 주장하였다. 그녀는 이 유형의 비용이 많이 드는 부가가치 색인의 생산을 사용료의 부과방법으로 충당할 수 있다고 제안하였다. Owen도[70] 전문적 색인작성이 인터넷을 통해 접근할 수 있는 고가치 DB를 위해 정당화 될 수 있다고 주장하였다. 확실히 자동색인과 초록 작성방법이 꾸준히 향상될 것이다. 그러나 Lancaster와 Smith가[71] 이 분야의 고찰에서 지적한 것처럼 기계가 이처럼 중요한 활동에서 완전히 인간을 대신할 만큼 충분히 지적이 되기에는 아마도 매우 오랜 시간이 걸릴 것이다.

V. 하이퍼텍스트/하이퍼미디어 링크

하이퍼텍스트와 하이퍼미디어 자원과 연관된 색인작성이 상당한 관심사로 등장하였다. 약간의 연구(Salton과 Buckley,[72]; Savoy,[73]; Salton 등,[74])가 하이퍼텍스트 링크를 자동으로 설정하는 방법을 검토하였으며, Agosti 등은[75] 실시간에 웹과의 상호작용을 브라우징하는 통계적 연관기준을 사용하여 하이퍼미디어 링크를 자동으로 설정하는 방법을 기술하였다. 그들은 이것을 하이퍼미디어의 "automatic authoring"이라고 하였다. 비록 그들이 이를 명백하게 주장하고 있지는 않지만 어떤 탐색자에 의해 설정된 하이퍼텍스트 링크는 다음 탐색자들에게 유용할 수 있다. 이 아이디어는 "성장하는 시소러스"(growing thesaurus)의 아이디어와 개념적으

69) Weinberg, B. H, "Complexity in indexing systems - abandonment and failure : implications for organizing the Internet," *Proceedings of the American Society for Information Science,* Vol. 33(1996), pp. 84-90.

70) Owen, P, "Structured for success : the continuing role of quality indexing in intelligent information retrieval systems," *Online Information,* Vol. 94(1994), pp. 227-231.

71) Lancaster, F. W. and Smith, L. C, *Intelligent Technologies in Library and Information Service Applications : a Realistic Appraisal.* Medford, NJ : Information Today, 1998.

72) Salton, G. and Buckley, c, "Automatic text structuring experiments," In : *Text-Based Intelligent Systems ; ed. by P. S. Jacobs,* pp. 199-210. Hillsdale, NJ: Lawrenec Erlbaum, 1992.

73) Savoy, J, "A new probabilistic scheme for information retrieval in hypertext," *New Review of Hypermedia and Multimedia,* Vol. 1(1995), pp. 107-134.

74) Salton, G. et al, "Automatic text structuring and summarization," *Information Processing & Management,* Vol. 33(1997), pp. 193-207.

75) Agosti, M. et al, "Automatic authoring and construction of hypermedia for information retrieval," *Multimedia Systems,* Vol. 3(1995), pp. 15-24.

로 유사하다.

Arents와 Bogaerts는[76] 하이퍼미디어 검색에 관한 문헌을 검토하였다. 비록 이들이 "색인작성"을 자주 언급하고 있지만 그들이 검토한 대다수의 방법(거의가 실험적인 방법)은 미리 설정된 링크나 또는 탐색과정 중에 형성된 링크를 따라가면서 하이퍼네트워크내에서 수행하는 브라우징이나 또는 "항해"(navigation)를 포함하고 있다. 이용자에게 네트워크내의 링크에 대한 시각적 개관을 제공하기 위해 설계된 그래프의 "브라우저"나 "지도"(maps)는 거의 40여년 전에 Doyle이[77] 제안한 "어의적 진로 지도"(semantic road maps)를 연상케 한다.[78]

Tessier는[79] 하이퍼텍스트 연결과 전통적인 색인작성 사이의 유사성을 검토하였는데, 그녀는 하이퍼텍스트 저자들은 텍스트를 전통적인 색인작성에서 연결하는 방법과 매우 유사한 방법으로 연결한다고 주장하였다.

Ellis 등은[80] 관습적인 색인자처럼 텍스트들의 집서에 하이퍼텍스트 링크를 삽입하도록 요청된 사람들이 이 업무에 많은 일관성을 보이지 못한 것을 발견하였다. Ellis 등은[81] 또한 검색효과에 대한 연결의 일관성의 영향을 시험하였다.

Chu는[82] 하이퍼텍스트 링크를 위해 망라성과 특정성의 원리를 적용하려고 시도하였다. 망라성의 측정은 정확할 수 있는 반면에(문헌내 단어수에 대한 링크 수), 특정성의 측정은 성공적으로 적용하기에 훨씬 더 어렵다.

Dimitroff와 Wolfram은[83] 정보검색활동에서 하이퍼텍스트DB의 사용과 연관된 문제들을 검토하였고, 그 뒤 Wolfram은[84] 세가지 다른 모델을 사용하여 하이퍼텍스트 레코드간 연결을 자세하게 조사하였다.

---

76) Arents, H. C. and Bogaerts, W. F. L, "Concept-based indexing and retrieval of hypermedia information," *Encyclopedia of Library and Information Science,* Vol. 58, Sup. 21(1996), pp. 1-29.

77) Doyle, L. B, "Semantic road maps for literature searchers," *Journal of the Association for Computing Machinery,* Vol. 8(1961), pp. 553-578.

78) Zizi, M, "Interactive dynamic maps for visualisation and retrieval from hypertext systems," In : *Information Retrieval and Hypertext ; ed. by M. Agosti and A. F. Smeaton,* pp. 203-224. Boston : Kluwer, 1996.

79) Tessier, J. A, "Hypertext linking as a model of expert indexing," *Advances in Classification Research,* Vol. 2(1992), pp. 171-178.

80) Ellis, D. et al, "On the creation of hypertext links in full-text documents : measurement of interlinker consistency," *Journal of Documentation,* Vol. 50(1994), pp. 67-98.

81) Ellis, D. et al, "On the creation of hypertext links in full-text documents : measurement of retrieval effectiveness," *Journal of the American Society for Information Science,* Vol. 47(1996), pp. 287-300.

82) Chu, H, "Hyperlinks : how well do they represent the intellectual content of digital collection?," *Proceedings of the American Society for Information Science,* Vol. 34(1997), pp. 361-368.

83) Dimitroff, A. and Wolfram, D, "Design issues in a hypertext-based information system for bibliographic retrieval," *Proceedings of the American Society for Information Science,* Vol. 30(1993), pp. 191-198.

84) Wolfram, D, "Inter-record linkage structure in a hypertext bibliographic retrieval system," *Journal of the American Society for Information Science,* Vol. 47(1996), pp. 765-774.

Ⅵ. 결 론

1) 오늘날 텍스트처리에서 사용되는 여러 가지 방법들은 특별히 새로운 것이 아니다. 대부분의 방법은 아마도 초보적 형태로 Luhn, Baxendale, Edmundson, Borko, Maron, Simmons, Salton 및 많은 연구자들에 의해 30여년 전에 이미 사용되었다.[85] 보다 훌륭한 결과가 오늘날 달성될 수 있는데, 이는 대량의 전자적 텍스트가 현재 이용가능하고 또한 현대의 컴퓨터 능력이 이러한 텍스트의 처리를 효율적으로 수행할 수 있기 때문이다.

그럼에도 불구하고 가장 정교한 현행의 방법조차도 달성된 결과, 처리시간 및 처리비용면에서 볼 때 이상적인 것과는 거리가 멀다. 더욱이 매일매일 진정한(real)서비스를 제공한다는 의미에서 실제로 운영되고 있는 시스템은 아직도 극히 소수에 불과하다.[86][87]

2) 현대의 가장 좋은 구문분석기도 일반적으로 비교적 짧고 단순한 문장만을 취급할 수 있을 뿐이다.[88] 보다 길고 복잡한 문장에 대해 분석기가 할 수 있는 최선은 구성요소의 단편(예컨대, 명사구)을 식별할 수 있을 뿐이며, 아직은 완전하고 모호하지 않은 분석을 할 수 있는 능력은 거의 없다.

3) 대부분의 자동색인은 이용자 중심적이라기 보다는 오히려 문헌-지향적이라고 볼 수 있다. 그러나 보다 이용자-중심적이라고 할 수 있는 방법 - 예컨대, 특별히 텍스트에서 탐색되어야 하는 용어들의 리스트를 사용하는 경우 - 이 개발되고 있다. 완전자동시스템은 자연어 질문, 적합성 피드백, 순위화된 출력 등을 제공해 줌으로써 출력면에서 보다 이용자-중심적이라고 볼 수 있다.[89]

4) 텍스트검색을 색인된 DB의 검색과 비교한 대부분의 연구들이 심각한 결함이 있다고 지적되었는데 이점은 불행하게도 자동색인과 수작업색인의 비교에서도 동일하다고 지적될 수 있다.[90]

Hmeidi 등은[91] "자동색인이 최소한 수작업색인만큼 효율적이며, 또 어떤 경우에는 보다 효

---

85) Lancaster, F. W, *Information Retrieval Systems : Characteristics, Testing and Evaluation.* New York : Wiley, 1968.

86) Jacobs, P. S, "Introduction : text power and intelligent systems," In : *Text-Based Intelligent Systems ; ed. by P. S. Jacobs,* pp. 1-8. Hillsdale, NJ: Lawrenec Erlbaum, 1992.

87) Lancaster, F. W. and Smith, L. C, 1998. op. cit.

88) McDonald, D. D, "Robust partial-parsing through incremental, multi-algorithm processing," In : *Text-Based Intelligent Systems ; ed. by P. S. Jacobs,* pp. 83-99. Hillsdale, NJ: Lawrenec Erlbaum, 1992.

89) Fidel, R, "User-centered indexing," *Journal of the American Society for Information Science,* Vol. 45(1994), pp. 572-576.

90) Lancaster, F. W. and Smith, L. C, 1998. op. cit.

91) Hmeide, I. et al, "Design and implementation of automatic indexing for information retrieval with Arabic documents," *Journal of the American Society for Information Science,* Vol. 48(1997), pp. 867-881.

율적이라고" 주장하고 있지만 이는 단지 컴퓨터분야의 아라비아어 초록의 소규모 DB의 검색 결과에 근거한 것으로 실제로는 수작업색인이 포함되지 않고 다만 Salton의 연구에 입각한 자동색인과 초록에 적용된 텍스트탐색과의 비교 결과에 근거한 주장이다. 그러나 Salton식 방법들이 재현율과 정도율의 향상(예컨대, 고빈도 단어들과 저빈도 단어들의 제거를 통해서)에 목적을 두고 있기 때문에, 이러한 방법들이 아마도 텍스트를 어간/어근 형태로 축소하는것 이외에는 아무것도 하지 못하는 것보다 더 좋은 결과를 달성한 것은 당연하다.

5) 컴퓨터의 사용비용이 계속 하락하는 사실에도 불구하고, 현대의 텍스트처리작업이 필연적으로 값이 싼 과제는 아니다. 현존하고 있는 비교적 업무가 한정된 시스템의 개발도 매우 비싼데, 예를 들면 CONSTRUE시스템의 경우 연간 9.5명의 인건비에 해당하는 비용이 든다.

6) 현대의 텍스트처리 시스템에 대해 보고된 50/50(재현율과 정도율)유형의 결과는 1960년대의 대규모 참조검색시스템(예컨대, MEDLARS)의 성능수준에 매우 가깝다.[92] 그러나 이 비교는 외관상으로 공정치 못한 것 같다. 왜냐하면 텍스트발췌 업무는 참조검색보다 분명히 더 복잡하기 때문에 보다 복잡한 업무에서 사용된 집서(corpora)는 비록 30여년전의 대규모 서지DB와 비교 하더래도(대략 1,500메시지 대 50만 서지레코드) 매우 적은 것으로 인정되어야 한다. 보다 전형적인 검색실험에서는 현대의 텍스트탐색방법들이 아주 대규모 DB(수십만 아이템)가 포함될 때에는 50/50 수준의 성능에 도달하지 못하였다.[93][94]

7) 오늘날 대규모 서지DB의 탐색을 위해 가장 일반적으로 사용되고 있는 비교적 조잡한 부울탐색방법들이, 많은 비판에도 불구하고, 취급된 집서(corpora)의 크기와 비교해서 괄목할 만큼 좋은 결과를 생산한다고 Stanfill과 Waltz는[95] 매우 설득력있게 다음과 같이 지적하였다.

"놀란만한 일은(인공지능의 관점에서 볼 때) 전혀 특정분야의 지식이 아닌 통계적 방법이 이용되고 있다는 사실과 또한 이 방법은 인공지능의 기준에 의하면 상상할 수 없는 대규모 정보량(기가바이트)에 사용되고 있다."

그러나 그들은 색인된 DB(예컨대, MEDLINE)나 또는 전문DB(예컨대, NEXIS)에서 사용하는 방법과 같은 단순한 부울탐색방법에 관련된 것이지 보다 정교한 순위화된 출력방법에 관련된 것이 아니라는 점을 주목해야 한다.

8) 오늘날 텍스트처리분야의 연구자들이 직면하고 있는 몇가지 도전은 시스템을 더욱 건전

---

92) Lancaster, F. W, *Evaluation of the MEDLARS Demand Search Service.* Bethesda, MD : National Library of Medicine, 1968.

93) Harman, D, "The TREC conferences," In : *Readings in Information Retrieval ; ed. by K. Sparck Jones and P. Willett,* pp, 247-256. San Francisco : Morgan Kaufmann, 1997.

94) Sparck Jones, K, "Reflections on TREC," *Information Processing & Management,* Vol. 31, No. 3(1995), pp. 291-314.

95) Stanfill, C. and Waltz, D. L, "Statistical methods, artificial intelligence, and information retrieval," In : *Text-Based Intelligent Systems ; ed. by P. S. Jacobs,* pp. 215-225. Hillsdale, NJ: Lawrenec Erlbaum, 1992.

하게 만드는 것(보다 훌륭한 정확성, 신속성, 언어분석 비용의 저렴성), 정제능력(예컨대, 문헌검색에서 문장(passage)검색으로 그리고 해답검색으로의 진전) 및 비용-효과면에서 더 좋은 그리고 이용자에게 더 매력적인 출력물의 작성(강조(highlighting), 텍스트발췌, 또는 요약) 등으로 확인되었다.[96]

9) 인터넷 정보자원을 보다 효율적으로 색인하고 탐색하기 위해 다양한 탐색엔진과 정보에이전트들이 개발되었다. 대부분의 탐색엔진이 1960년대 초의 전형적인 도치파일의 구축방법이지만 규모면에서 엄청나게 큰 도치파일의 효과적 갱신과 탐색을 허용하고 탐색방법도 보다 세련되고 있다. 상당한 에이전트가 유용한 업무를 수행하고 비교적 정교하다고 판정될 수 있지만 대부분의 에이전트가 비교적 단순한 텍스트탐색이나 또는 키워드 발췌방법에 주로 의존하고 있어 진정으로 지적 에이전트라고 하기에는 미흡하다. 그러나 새로운 도구와 성능이 거의 매일 추가되고 있는 사실로 미루어 자동색인 및 초록 작성방법이 꾸준히 향상될 것이다.

10) 정보검색에 관련된 다양한 업무를 위한 컴퓨터의 적용이 상당히 발전하였지만 자동방법들이 색인작성, 초록작성, 시소러스 구축 및 탐색전략의 고안과 같은 지적업무에서 조만간 인간보다 성능이 우수할 것이라는 징후는 거의 없다.

〈참고문헌은 각주로 대신함〉

96) Jacobs, P. S. "Introduction : text power and intelligent systems," In : *Text-Based Intelligent Systems ; ed. by P. S. Jacobs*, pp. 1-8. Hillsdale, NJ: Lawrenec Erlbaum, 1992.