情報檢索分野에 미치는 人工知能의 影響

The Influence of Artificial Intelligence on the Information Retrieval System

金 泳 煥*
(Kim, Young-Whan)

抄 錄

情報檢索分野의 研究課題들을 구분하여 정리하였으며 人工知能의 重要研究分野를 소개하고 이들이 情報檢索에 응용될 수 있는 특성에 대해서 살펴보았다. 그리고 情報檢索分野에 이용될 수 있는 人工知能의 應用分野中 현재까지 활발히 연구되고 있는 몇가지 分野를 살펴봄으로써 情報檢索에서의 人工知能技術의 應用可能性과 그 影響을 정리하였다.

ABSTRACT

The definition of information retrieval and artificial intelligence is given and the research activity in information retrieval, as well as the major artificial intelligence techniques which can be applied to information retrieval problems, is reviewed. By outlining the several artificial intelligence application in information retrieval, the potential role of artificial intelligence in information retrieval is discussed.

* 韓國科學技術院 電算學科 博士課程 在學中(人工知能 專攻).
現韓國電氣通信公社 事業支援團 專任研究員.

Ⅰ. 序　論

人工知能과 情報檢索(Information Retrieval : IR)은 각기 다른 필요성에 의하여 생겨난 분야이며 그 추구하는 目的과 發生動機도 각기 다르다.

人工知能에 대한 研究는 1960년대 말경부터 본격적으로 시작되었으며 사람이 소유하고 있는 知能(Intelligence)을 컴퓨터가 가질 수 있도록 하는 것이 窮極的인 目的이다. 즉, 言語를 이해하고, 學習能力을 가지고, 추론을 하며, 問題를 해결하는 것과 같은 人間이 가진 知能을 가지고 이러한 일들을 할 수 있는 知能型 컴퓨터시스템(Intelligent Computer System)을 설계하는 것이 그 目的이다.

초기에는 여러가지 問題들을 모두 해결할 수 있는 一般的인 問題解決方法(General Problem Solver)을 찾으려고 노력하였으나 이러한 일들이 매우 어렵고 實效性이 희박하다는 결론을 내리고 1970년대에 들어와서는 問題를 쉽게 해결하기 위한 表現方法과 빨리 해답을 얻을 수 있는 探索方法에 대한 研究가 집중적으로 이루어졌다. 그러나 이러한 방법에도 한계가 있음을 깨닫고 1970년대 말에는 문제에 내포되어 있는 知識이 問題解決에 아주 큰 비중을 차지한다고 생각하여 좁은 영역(Domain)에서 그 領域에 속하는 지식을 활용하여 問題를 해결하고자 하는 專門家시스템 또는 知識基盤시스템에 대한 研究가 활기를 띠기 시작하였다. 이 외에도 컴퓨터시각, 자연언어처리, 문제해결, 계획 등의 여러 應用分野가 있다.

科學技術의 급격한 발달로 인하여 情報의 量이 폭발적으로 증가함에 따라 사용자들에게 그들의 요구에 맞는 適切한 最新의 情報를 제공하여 주는 일이 아주 어려운 문제로 대두되었다. 이러한 연유로 1960년대 초에 컴퓨터를 이용하여 이러한 문제들을 해결하려는 것이 情報檢索시스템의 生成動機이다. 情報檢索시스템은 사용자의 요구에 가장 적합한 정보들을 제공하여 주는 것이 그 목적이다. 情報檢索의 主要研究課題는 정보의 표현, 저장, 구성, 검색에 관한 것이다. 이러한 情報檢索시스템은 다루는 情報의 形態에 따라서 크게 데이터베이스 管理시스템(Database Management System), 文書檢索시스템(Document Retrieval System, Reference Retrieval System), 自動質疑應答시스템(Automatic

Question Answering System)으로 나눌 수 있는데 여기에서는 주로 文書檢索 시스템에 대해서 이야기하고자 한다.

이와 같이 각기 다른 目的으로 두 分野가 발전되어 왔지만 궁극적으로 볼 때 두 분야 모두가 컴퓨터를 問題解決의 手段으로 본다는데 공통점이 있다. 결국 한 分野에서 해결하지 못한 問題點들을 다른 분야에서 開發된 技法과 槪念들을 이용하여 해결할 수 있다. 특히 情報檢索分野에서는 다른 분야보다도 훨씬 많이 人工知能의 影響을 받아왔으며 앞으로도 人工知能의 技法과 槪念들을 계속해서 응용할 수 있는 분야이다.

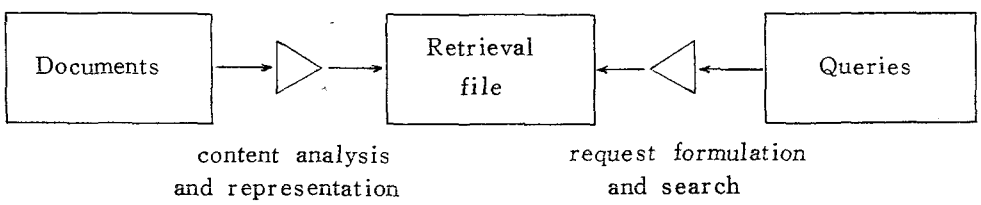
이와 같은 연유로 인하여 人工知能이 情報檢索分野에 미치는 影響을 分析하여 보고 현재 해결하여야 할 問題點들과 그 문제들의 解決方案에 대해서 살펴보려고 한다. 그리고 人工知能의 技法을 이용하여 개발된 情報檢索시스템의 여러 特性들을 분석하고 앞으로의 硏究方向을 살펴보려고 한다.

Ⅱ章에서는 情報檢索시스템에 대한 기본적인 개념들을 소개하고 현재의 硏究內容과 앞으로의 發展方向을 정리한다. Ⅲ章에서는 情報檢索시스템에 미치는 人工知能의 影響을 人工知能의 각 분야별로 간략히 살펴보고 현재 주로 많이 연구되는 '情報檢索에서의 人工知能 應用分野'에 대해서 조사한다. 그리고 情報檢索에 人工知能의 技法을 이용한 한 예를 살펴본다. Ⅳ章 結論에서는 앞으로의 硏究方向과 內容에 대해서 언급한다.

Ⅱ. 情報檢索시스템

情報檢索이란 情報의 構造, 分析, 組織, 貯藏과 情報의 檢索에 관한 硏究分野로서 사용자의 요구에 가장 적절한 정보를 찾아내는 것이 목적이다.

情報檢索시스템의 主要構成要素는 다음과 같다.



관련된 문서들을 정해진 分析方法을 이용하여 분석한 후 이것을 파일에 저장하고 사용자가 質疑語(Query)를 이용하여 요구하는 내용을 표현하면 이를 분석하여 시스템에 定義된 形態로 바꾼 뒤에 質疑語의 內容을 가지고 문서들 중에서 가장 적절한 文書를 찾게 된다. 기존의 情報檢索시스템은 크게 데이터베이스 管理시스템, 文書檢索시스템, 自動質疑應答시스템으로 분류할 수 있다. 이 세 가지 시스템들은 周邊컴퓨터의 技術發達로 인하여 배치 시스템에서 온라인 시스템으로 발달되어 왔으며 이로 인하여 使用者가 즉시 원하는 정보들을 얻을 수 있게 되었으며 또한 얻은 정보들을 이용하여 반복해서 여러번 검색을 할 수 있게 되었다. 이와 같이 함으로써 처음에 원하는 내용을 質疑語로 표현하여 검색을 한 후 檢索된 情報를 이용하여 質疑語를 수정, 보완하여 검색을 반복함으로써 적절한 情報를 유출할 수 있다. 또한 한 사용자가 여러개의 情報源을 동시에 검색할 수 있으며 여러 使用者가 형성된 情報네트워크를 통하여 하나의 정보원을 공용할 수 있다. 물론 이와 같은 네트워크로 인하여 情報의 Security 문제, 資料의 Integrity 문제, Privacy 문제 등이 대두되었다. 여기에서는 3가지 시스템 중 주로 文書檢索시스템에 대해서 고려하기로 한다.

文書檢索시스템은 사용자가 원하는 문서를 찾아주는 것이 그 목적이며 文書의 內容을 분석하여 이것을 索引構造로 구성을 하고 사용자의 質疑語에서 사용자가 요구하는 情報의 內容을 명시하도록 하여 文書의 集合에서 정해진 探索方法에 의해 적절한(relevant) 문서를 찾아준다. 이 시스템에서 다루는 주요한 문제들은 다음과 같다.

① 自動索引技法(Automatic Indexing Technique)
② 言語自動標準化技法(Automatic Language Normalization Technique)
③ 自動分類方法(Classification)
④ 檢索方法
⑤ 質疑語 自動形成方法

이와 같은 시스템들을 설계할 때는 사용자가 쉽고 편리하게 시스템을 사용할 수 있도록 해야만 한다. 기존의 시스템들은 시스템을 效果的으로 사용하기 위하여 그 시스템에 대해서 숙달된 사용자이어야만 원하는 情報를 쉽게 얻을 수 있으며, 使用者가 원하지 않는 情報들을 쓸데없이 供給하는 短點이 있다. 이러한 관점에서 볼 때 基本的으로 시스템이 갖추어야 할 사항들은 다음과 같다.

① 專門知識이 없는 사용자도 쉽게 시스템을 사용할 수 있는 使用者 인터페이스의 設計

② 情報에 대한 使用者와 시스템間의 관점의 차이를 줄일 수 있는 融通性있는 分類시스템(Classification System)과 다양한 情報接近技法

③ 저장해야 할 情報파일의 크기를 될 수 있는 한 줄일 수 있는 情報比較法과 文書의 抽象化(Text Abstraction)方法

④ 불필요한 情報들을 빨리 걸러낼 수 있는 빠른 Text Scanning과 Document Skimming方法

또한 使用者 質疑語를 자동으로 Boolean Query로 바꿀 수 있어야 한다. Boolean Query는 주로 Single Term과 anded term clause들의 disjunctive normal form으로 표현된다. 그 表現形態는 다음과 같다.

$$T_1 \text{ or } T_2 \text{ or } \cdots\cdots \text{ or } T_m \text{ or } (T_i \text{ and } T_j) \text{ or } (T_k \text{ and } T_n) \text{ or } \cdots\cdots \text{ or } (T_n \text{ and } T_p \text{ and } T_q)$$

각각의 term들은 文書集合內에서 나타나는 빈도수에 따라서 분류되며 빈도수가 적은 term들은 Weight가 높으며 주로 Single term으로 사용되고 빈도수가 큰 term들은 Weight가 낮으며 주로 anded term clause로 사용된다. 이와 같이 Boolean query를 구성하여 만족하는 情報要求가 나타낼 때까지 반복적으로 수정하여 나간다. 현재의 시스템들은 一般大衆使用者들이 쉽게 사용할 수 있는 시스템이 되기 위해선 해결해야 할 問題點들이 아직도 많이 있다. 使用者인터페이스의 問題가 아직 완전히 해결되지 못한 대신 情報分析과 檢索方法에는 그동안 많은 발전이 이루어졌다. 그것은 크게 볼 때 自動索引, 自動檢索技法, 質疑語 自動生成으로 나눌 수 있다.

自動索引은 文書의 內容을 분석하여 그 文書의 內容을 잘 나타낼 수 있는 索引(Index, Keyword)들을 뽑아내는 것이며, 여러가지의 기법들이 있다. Full-text indexing system의 경우에는 기능어들을 모아둔 "Stop-list"에 나타나지 않는 모든 단어가 인덱스가 되며 Word indexing system에서는 일부 단어들만이 인덱스가 된다. 의미가 너무 넓은 단어들(빈도수가 많은 단어)은 관용어구 사전을 이용하여 좀 더 具體的인 意味를 가진 관용어구로 대치하고 의미가 너무 좁은(빈도수가 적은 단어) 단어들은 같은 의미를 가진 단어들을 한데 묶

어 놓은 辭典(Thesaurus)을 이용하여 같은 의미를 지니면서 좀 더 一般的인 單語로 대치한다. 또한 Cluster identifier를 이용한 Cluster indexing system도 사용할 수 있다. 自動索引過程의 一般的인 順序는 다음과 같다.

① 문서들 중에 나타나는 각각의 단어를 lexical analysis를 통하여 뽑아낸다.

② 機能語(and, of, but, the 등)들을 모아둔 'Stop-list'를 이용하여 機能語를 제외시킨다.

③ 接尾辭 分離過程(suffix stripping routine)을 이용하여 접사부분을 없앤 어근(word stem)을 찾아낸다.

④ 각 단어에 대하여 Weight를 계산한다(inverse document frequency weight). 즉 빈도수가 적은 單語가 높은 weight를 가진다.

⑤ 아주 頻度數가 많거나, 적은 단어들을 구별하기 위한 경계치(frequency threshold)를 정한다.

⑥ 아주 頻度數가 적은 단어들은 같은 의미를 지닌 단어들을 한데 묶어 놓은 辭典(Thesaurus)을 이용하여 같은 의미를 지니면서 좀 더 一般的인 의미의 단어로 대치한다.

⑦ 아주 頻度數가 많은 단어들은 관용어구 사전(Phrase dictionary)을 이용하여 좀 더 具體的인 意味를 가진 관용어구로 대치한다.

⑧ 각 文書들을 지금까지 찾아낸 단어와 그의 Weight의 쌍으로 표현한다.

⑨ 使用者 質疑語를 각 문서들의 Weight term set들과 비교하여 가장 적합한 文書를 골라낸다.

⑩ 檢索된 文書가 사용자가 요구하는 內容에 적합한 것인지 아닌지를 구분한다.

⑪ term relevance weight (적합도)를 계산한다.

⑫ 원래의 使用者 質疑語와 적합하다고 구분된 文書에서 얻은 단어들을 이용하여 改善된 使用者 質疑語를 재구성하고 각 단어마다 inverse document frequency와 relevance를 고려하여 weight를 부여한다.

⑬ ⑨단계로 가서 使用者가 만족하는 文書를 얻을 때까지 반복한다.

自動檢索技法은 요구하는 정보를 표현한 質疑語의 內容을 문서집합들과 비교하는 것으로서 아주 單純한 方法으로는 順次的 檢索이 있다. 이 方法은 하나하나씩 처음부터 끝까지 비교하는 方法으로 간단하고 완전한 방법이지만 시간이

많이 걸리는 短點이 있다. 대신 적은 양의 文書集合에 대해서는 적합한 방법이다. 또 한 가지 방법으로는 副索引(auxiliary index)을 사용해서 inverted file을 구성하여 檢索하는 方法이 있다. 그 외에도 문서들을 cluster 하여 clustered file을 구성하여 cluster tree-search를 하는 방법도 있다. 이 方法에서는 文書를 cluster 하기가 어렵다는 問題點이 있다.

質疑語 自動生成은 사용자가 표현한 質疑語를 시스템이 구조에 맞도록 再構成하는 과정으로서 單純한 方法으로는 Boolean query 方法이 있다. 이 方法에서는 우선 Boolean query를 구성하기가 힘들고 weight를 사용할 수 없고 檢索된 資料들을 適合性程度(measure of relevance)에 따라 순위를 매길 수가 없다. 특히 or나 and의 사용으로 인해 문서들간의 適合性程度의 比較가 불가능하다. 이와 같은 問題點들을 해결하기 위하여 一般化되고 改善된 Boolean Query 方法을 사용하고 있다. 이 중의 하나가 fuzzy set model 이며 이 方法은 일반적인 Boolean query 方法과 같고 대신 文書內의 單語에 대해서는 weight를 부여할 수 있는 方法이다. 하지만 이 方法은 Boolean query 方法에서 발생한 or, and로 인한 問題點들을 해결하지 못한다.

또 하나의 方法은 " Extended Boolean Query System "으로서 Boolean operator의 적용을 融通性있게 한 방법이다. 이 方法의 基本的인 方針은 質疑語에 나타난 단어를 더 많이 포함하고 있는 문서가 그렇지 못한 文書보다 더 가치있는 것으로 취급된다는 사실이다. 이 方法에서는 一般化된 Distance Function을 사용하는데 이 함수는 p라는 변수를 가지며 이 p는 그 값이 1에서 무한대 사이에 존재한다. 만약 p가 무한대이면 일반적인 Boolean Query에서와 같이 Boolean operator가 그대로 적용된다. p가 무한대에서 1로 가까이 갈수록 Boolean operator and나 or가 원래의 의미를 점점 잃게 된다. 즉, and operator는 모두 만족하는 것에서 대부분이 만족하면 되는 것으로 그 의미가 바뀌게 되고 or operator는 하나만 만족하면 되는 것에서 여러 개가 만족하면 되는 것으로 그 의미가 바뀌게 된다.

p가 1인 경우에는 and, or operator간에는 아무런 차이가 없이 단순한 vector query로 이해된다.

이와 같은 方法으로 Extended Boolean query system은 일반적인 query structure를 그대로 사용하면서 훨씬 좋은 similarity function을 가질 수 있으

며 文書와 query에 weight를 부여할 수 있으며 檢索된 文書들에 rank를 부여할 수 있다. 이 시스템은 實驗結果 일반적인 Boolean query system보다 훨씬 向上된 性能을 가짐이 판명되었다.

Extended Boolean Retrieval System의 基本的인 處理過程은 다음과 같다.

① 적절한 文書의 대부분을 모두 檢索할 수 있도록 넓은 의미를 지닌 일반적인 Boolean Query를 작성한다.

② Inverted File System을 이용하여 ①에서 작성한 質疑語로 적절한 文書들(D)을 고른다.

③ ①에서 작성한 質疑語를 이용하여 Extended Boolean System에서 p값이 1에서 3사이인 質疑語를 作成한다.

④ ③에서 만들어진 質疑語를 ②에서 고른 文書集合 D에 적용하여 각 문서들을 質疑語와의 similarity정도에 따라 순서를 부여한다.

⑤ ④에서 골라진 적절한 文書의 單語들을 이용하여 改善된 質疑語를 재구성하여 ④단계로 가서 使用者가 만족하는 결과를 얻을 때까지 반복한다.

기존의 情報檢索시스템의 性能을 평가하는 기준은 Precision과 Recall이 있다. 이 두 가지 기준의 정의는 다음과 같다.

$$\text{Precision} = \frac{\text{number of relevant items retrieved}}{\text{number of items retrieved}}$$

$$\text{Recall} = \frac{\text{number of relevant items retrieved}}{\text{number of relevant items in the files}}$$

Precision은 檢索된 情報들 중에 사용자가 요구하는 情報가 어느 정도 포함되어 있는지를 나타내는 값으로서 시스템의 精密度를 測定한다.

Recall은 全體文書 集合中에서 사용자가 요구하는 문서들을 어느 만큼 많이 찾아낼 수 있는가를 나타내는 값이다. 이 두 기준은 서로 반비례적인 상관관계를 갖고 있다.

시스템이 점점 發展될수록 이러한 기준 외에 다른 性能評價基準이 필요하고 이에 대한 硏究도 진행되고 있다.

Ⅲ. 人工知能의 影響

情報檢索分野에서는 벌써부터 人工知能의 技法들을 많이 사용하고 있다. 그럼에도 불구하고 人工知能의 技法들을 이용하여 개선할 수 있는 부분이 아직도 많다. 이 章에서는 人工知能이 情報檢索分野에 미칠 수 있는 影響에 대해서 살펴보기로 한다. 우선 人工知能의 各分野別로 어떻게 情報檢索分野에 影響을 미칠 수 있는가를 간단히 살펴보고, 情報檢索分野에 影響을 미칠 수 있는가를 간단히 살펴보고, 情報檢索分野에 適用할 수 있는 人工知能의 應用分野를 크게 네 가지로 나누어 살펴본 후 하나의 응용예에 대하여 좀 더 상세히 어떤 問題點들이 있는가를 살펴보기로 한다.

情報檢索시스템에 영향을 미치는 人工知能의 分野는 패턴認識, 知識表現, 問題解決과 計劃, 휴리스틱스, 學習 등으로 나눌 수 있다.

패턴認識은 주어진 데이터로부터 유용한 정보를 얻어내는 問題에 관한 分野로서 크게 두 개의 槪念이 있다. 즉, 入力資料에서 그 자료의 특성을 뽑아내는 特性抽出과 이 資料特性을 이용하여 미리 정의된 class 중 어떤 class에 속하는가를 결정하는 패턴分類(Pattern classification)가 있다. 特性抽出은 情報檢索시스템에서 주어진 문서들의 내용을 분석하여 그 문서가 의미하는 정보를 잘 표현할 수 있도록 해주는 과정인 內容分析(Content analysis)部分에 적용될 수 있고, 패턴分類는 시스템이 가지고 있는 情報들(文書集合) 중에서 사용자가 요구하는 情報를 찾아내는 matching process 部分에 적용할 수 있다.

知識表現은 人工知能시스템에 필수적인 요소인 지식을 어떻게 형식화하여 문제를 해결할 때 사용할 수 있는가를 다루는 分野로서 知識表現方法으로는 述語論理, 知識表現 프로그래밍言語, 規則基盤시스템, 意味網(Semantic Network), 프레임(Frames) 등의 여러 방법들이 있다. 情報檢索시스템의 知識表現問題는 각 문서로부터 뽑아낸 특성들을 어떻게 데이터베이스內에 표현하는가 하는 문제이다. 즉, 각 문서들의 內容과 意味를 가장 잘 표현할 수 있도록 각각의 文書를 표현하여야 한다. 索引過程에서 사용하는 Thesaurus (같은 의미를 가진 단어들을 한데 묶어놓은 辭典)는 人工知能에서 사용하는 의미망과 흡사하며 이 分野의 技術을 이용하여 그 구성을 효과적으로 할 수 있을 것이다. 質疑-應答

시스템의 경우에는 述語論理를 쓰면 아주 효과적인 방법이 된다. 어쨌든 情報檢索시스템이 일반적이고, 사용자가 쉽게 사용할 수 있는 性能이 우수한 시스템이 되기 위해서는 人工知能에서 사용하는 知識表現方法을 알맞게 응용하면 큰 도움이 될 것이다.

問題解決과 計劃은 주어진 문제를 해결하기 위하여 미리 알고 있는 지식을 이용하여 問題解決을 시도하고 이러한 과정에서 얻은 情報들을 이용하여 만족한 결과를 얻을 때까지 반복적으로 問題解決을 시도한다. 이 경우 主關心分野는 探索方法이 되며 問題解決을 위하여 가능한 방법들 중에서 어떤 것을 이용할 것인가를 다루는 計劃의 問題도 포함된다.

問題解決은 情報檢索시스템에서 볼 때 使用者質疑語가 하나의 문제이고 어떤 문서가 사용자요구에 가장 적절한 것인지를 결정하는 과정이 問題解決過程으로 볼 수 있다. 따라서 問題解決에서 사용하는 여러 방법들을 이용하여 사용자 요구에 가장 적합한 문서를 찾을 수 있는 探索方法을 개선할 수 있을 것이다. 또한 計劃技法을 이용하여 情報檢索시스템에서 경우에 따라 적합한 探索方法을 선택하는데 이용할 수 있다.

휴리스틱스는 잘 정의된 알고리즘이 존재하지 않는 問題를 해결할 때 사용될 수 있는 經驗的 知識을 이용하는 분야로서 情報檢索시스템의 경우에는 이 시스템使用의 專門家(Intermediary)가 경험적 지식을 이용하여 일반 사용자의 요구사항을 파악하여 잘 정의된 質疑語의 形成을 도와준다. 이 경우 專門家의 役割을 대신할 수 있는 專門家시스템을 만드는데 응용될 수 있다.

또한 人工知能의 學習槪念을 이용하여 學習機能을 가진 檢索시스템을 設計할 수 있다. 檢索시스템의 學習은 크게 두 가지 방법으로 나눌 수 있다. 그 하나는 데이터베이스를 探索하는 過程에서 얻은 정보를 이용하여 質疑語를 修正補完하는 " Relevance Feedback " 이고 또 하나는 質疑語에 적합하다고 판단되는 문서의 Document Vector 를 Query Vector 들의 內容을 이용하여 이와 비슷하게 내용을 수정하는 " Document Vector Modification " 이 있다.

지금까지 人工知能의 각 분야가 어떻게 情報檢索分野에 影響을 미치고 응용될 수 있는가를 간단히 살펴보았다. 그러면 情報檢索分野에 이용될 수 있는 人工知能의 應用分野를 다음과 같이 크게 네 분야로 나누어 살펴보기로 한다.

① Human-Database Interface

② Conceptual Indexing

③ Automatic Data Entry

④ Active Memory Techniques

이러한 응용분야들은 아직 實用化된 상태가 아니고 연구를 수행중인 實驗的인 狀態이다.

Human-Database Interface는 숙달되지 못하고 專門知識이 없는 사용자가 쉽게 컴퓨터시스템을 사용하도록 하기 위하여 使用者와 데이터베이스 사이에 人工知能 인터페이스를 開發하는 것이다. 自然語處理分野가 여기에 해당된다. 이 分野의 目的은 사용자가 아주 숙달된 人間專門家와 대화하면서 원하는 情報를 얻어내는 것과 같은 효과를 시스템에서도 할 수 있도록 하는 것이다. 이러한 목적을 달성하기 위해서 가장 必要한 것은 自然語를 處理, 理解할 수 있는 능력이다. 즉, 使用者가 원하는 要求事項을 자연언어로 표현하여 시스템과 대화를 나누면서 점차적으로 원하는 內容을 시스템에게 정확히 전달할 수 있는 능력을 말한다. 이러한 自然言語處理시스템은 기존의 情報檢索시스템의 前處理시스템(front-end) 역할을 하게 된다. 하지만 自然言語處理시스템을 설계하는 일은 매우 어려운 일이며 단순한 質疑語 프로세싱과 달리 해결하여야 할 여러가지 問題點들이 있다.

이러한 問題點들을 정리하여 보면 다음의 여섯 가지로 볼 수 있다.

① 自然語의 文法上 誤謬를 處理하는 問題

自然語處理시스템의 첫번째 단계는 syntactic parsing부분이다. 따라서 사용자가 문법적으로 잘못된 자연어를 사용하였을 때는 보통사람들이 대화할 때와 같은 意味傳達을 할 수 없게 된다.

② 代名詞의 處理問題(Pronoun Referents)

일상생활에서 사용되는 自然語에서는 대명사를 많이 사용하며 이러한 代名詞는 그 문맥을 이해함으로써 그것이 지칭하는 것이 무엇인지를 알 수 있다. 이러한 代名詞處理는 단순히 문법적인 분석으로는 해결하기가 힘들다.

③ Lexical Ambiguity

自然語에서 대부분의 단어는 하나의 의미만을 가지고 있지 않고 여러가지의 의미를 가지고 있다. 그 여러가지 의미 중에서 사용자가 의도하는 것이 무엇인지를 알아내어야만 한다.

④ 接續詞의 問題

接續詞를 이용한 문장에서 접속사로 연결된 단어들이 의미상으로 어떻게 처리되어야 하는가를 결정하여야 한다. 예를 들어 " Is it time to re-order the high-voltage diodes and transistor ? "와 문장에서 high-voltage가 transistor에도 적용되는지, 아닌지를 구분하여야 한다.

⑤ 修飾語句의 處理問題

修飾語句가 과연 어떤 단어를 수식하는지를 구분하여야 한다.

⑥ 文法的으로는 맞지만 의미상으로 틀린 文章의 處理問題

이러한 여러가지 問題點들을 해결하기 위해서 기본적인 방법은 실제 세계의 지식을 이용하여야 하고 文法的인 分析보다는 意味分析에 의존을 해야만 한다.

Conceptual Indexing은 데이터베이스를 그 구성원소의 의미를 나타내도록 구성하기 위한 것이다. 이 경우 質疑語도 그 의미를 잘 나타낼 수 있도록 구성되어야 한다. 만약 이와 같은 Conceptual Indexing이 성공하게 되면 기존의 시스템에서와 같이 使用者의 要求事項에 근사한 質疑語가 아닌 使用者가 의도하는 의미를 정확하게 표현하는 質疑語를 구성할 수 있으며 데이터베이스內에서 정확하게 필요한 것들만 찾아낼 수 있게 된다. 물론 이와 같은 方法에서는 索引過程 자체가 어려운 작업이 되므로 사용자가 이것을 할 수 없을지도 모른다. 그러한 경우에는 Human-Database Interface分野에서 知的인 使用者 인터페이스를 잘 설계함으로써 해결할 수 있다.

Conceptual Indexing에 관한 研究는 心理的인 側面에서부터 발전되어 왔다. 즉, 인간은 어떤 사실에 대한 質問을 받았을 때 자기가 소유하고 있는 情報들을 이용하여 굉장히 빨리 이에 대한 답을 내리는 능력이 있다. 이것은 현재 論理的으로 그 과정을 모델化할 수 없는 상태다. 만약 인간이 소유하고 있는 情報들을 어떻게 저장하고 관리하고 있는가를 알 수 있다면 이러한 方法을 情報檢索시스템에 적용하면 현재의 시스템보다 훨씬 우수한 시스템을 만들 수 있을 것이다.

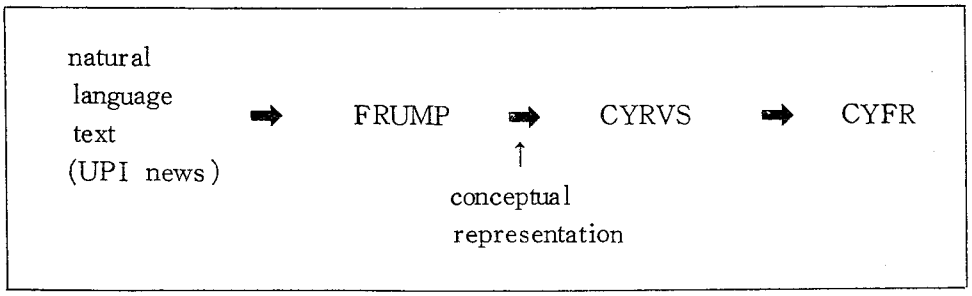
人間의 記憶構造를 이해하여 그 방법을 이용해보려는 시도로 만들어진 시스템이 " IPP "와 " CYRUS "이다. 이 두 시스템은 모두 Yale大學校의 Schank의 지도를 받은 학생들의 博士學位論文으로서 學習能力을 필수 불가결한 요소로 보고 있다.

IPP는 테러에 관한 뉴스記事를 기억하고 이에 대한 情報를 알려주는 시스템으로서 어떤 새로운 內容의 記事가 들어오면 미리 저장된 內容과 比較를 통하여 이 記事內容을 一般化시킨다. 이렇게 함으로써 그 기사의 개념적 의미를 메모리에 저장하게 된다.

CYRUS는 前美國務長官이던 Cyrus Vance의 職業生活을 모델로 한 것으로서 Vance에 대한 배경지식을 미리 가지고서 Vance의 새로운 행동에 대한 소식을 Conceptual Indexing方法을 사용하여 시스템내에 구성한다. 현재 이 분야에 대한 硏究는 Roger Schank와 Don Norman을 중심으로 활발히 이루어지고 있다.

Automatic Data Entry分野는 文書檢索시스템에 사용하는 데이터베이스는 수시로 변경되어야 한다. 즉 새로운 內容이 첨가되거나 어떤 경우에는 기존의 內容이 삭제되어야 한다. 기존의 데이터베이스 構成方法을 사용하는 시스템의 경우에는 이 작업을 하는데 시간이 많이 소모되지만 比較的 簡單한 作業이다. 하지만 데이터베이스를 Conceptual Indexing分野에서 논의한 것 같이 개념적으로 구성(conceptually organization)하였다면 데이터베이스의 內容變更이 쉬운 일이 아니다. 이 경우에는 새로운 내용을 데이터베이스에 첨가하려고 할 때 우선 그 內容에 대한 意味表現(meaning representation)이 선행되어야 한다. 물론 이러한 意味表現過程을 사람이 수행할 수 있지만 이렇게 되면 시간이 많이 걸리고, 실수할 경우가 생기고 또한 不一致問題가 발생하게 된다. 따라서 解決策은 人工知能시스템이 이 작업을 하도록 하는 것이다. 이러한 시스템에서 가장 중요한 것은 自然語로된 文書를 이해하여 시스템에서 사용하는 意味表現으로 바꾸어주는 작업이다. 이와 같은 것에 대한 硏究分野가 Automatic Data Entry이다. Automatic Data Entry와 Conceptual Organization of Memory의 개념을 결합하여 구성된 시스템이 CYFR(schank)이다.

CYFR의 構成은 다음과 같다.



CYFR은 FRUMP시스템과 앞에서 언급한 CYRVS를 결합한 시스템이다. FRUMP는 UPI의 뉴스를 입력으로 받아들여 보통사람이 新聞을 빠른 속도로 훑어볼 때 가질 수 있는 정도의 이해력을 가지고 각각의 뉴스를 63개의 부류로 나누게 된다. 그 다음 이것을 槪念的 表現으로 바꾸고 그 중에서 Vance에 관련되는 모든 표현을 CYRVS시스템으로 넘겨주게 되고 CYRVS시스템은 이 내용을 메모리에 저장한다. 이 CYFR은 현재 硏究用으로 이용되고 있다.

Active memory technique分野란 Active memory란 情報를 檢索하고 수정할 때 메모리 자체가 能動的인 役割을 하는 시스템을 말한다. 사람을 예로 든다면, 만약 사람이 어떤 사실에 대해서 질문을 받았을 때 그 사실 자체에 대한 情報가 없더라도 관련된 다른 情報들을 이용하여 그에 해당하는 情報를 만들어낼 수 있다. 이와 같은 메모리를 " recenstructive "하다고 표현한다. 물론 이와 같은 일이 情報檢索의 많은 경우에 사람이 항상 실패없이 할 수 있는 것은 아니지만 이러한 개념을 이용한다면 데이터들을 일일이 기억하고 있을 필요없이 그 데이터들을 특정지우는 法則을 가지게 되면 똑같은 효과를 얻을 수 있을 것이다. 이와 같은 active memory技法을 이용하게 되면 데이터베이스 一致性檢査나 推論能力을 데이터베이스 자체가 가지게 된다. 이것을 실현시키는 方法을 데이터베이스 시스템내에 整理證明機를 두는 것이다. 즉 質疑語가 들어왔을 때 일단 적합한 해답을 찾고 만약 찾지 못했을 때는 데이터베이스내에 있는 情報를 근거로 하여 質疑語를 증명하면 된다. 이와 같은 方法을 사용하는 것이 " Deductive data retrieval "이다.

Deductive data retrieval system은 현재까지 많이 개발되었으며 주로 일차술어논리를 사용하여 개발되었다. 또 다른 問題는 " monotonicity "이다. 즉 어떤 정리가 한번 증명이 되면 그 후에 더 많은 情報의 獲得으로 인하여 그 정리가 거짓이 판명되었는데도 이를 증명할 수 없게 된다. 이로 인하여 정보들 사이에 不一致問題가 발생하게 된다. 실제로 사람의 행위를 살펴볼 때 이러한 monotonicity는 발생하지 않는다. 사람은 참이라고 가정했던 어떤 사실이 거짓이라고 판명되면 그 가정을 포기하고 그 가정으로 인해 확신했던 다른 사람들도 포기하게 된다. 이와 같이 하기 위해서는 비단조논리(non-monotonic logic)를 사용하여야만 한다. 이와 관련된 연구로서 데이터關聯性(data dependency) 연구가 있으며 이는 데이터베이스의 一致性問題를 해결하는데 응용할 수 있다.

Active memory와 관련된 또 다른 분야로 " Inductive inference "分野가 있다. Inductive inference는 入力資料들의 集合에 대해서 특성을 분석하여 그 개념을 一般化 (generalization)하는 것이다. 이와 관련된 주된 分野는 學習과 推論이다. 하지만 이것은 情報檢索에 응용될 수 있으며 情報檢索시스템의 중요한 요소가 될 것이다.

위에서 살펴본 바와 같이 情報檢索시스템의 많은 부분이 人工知能의 影響을 받아 그 기법을 이용하여 개선되고 있다. 그 예로서 知識基盤시스템 接近方法을 들 수 있다.

실생활에서 살펴볼 때 효과적인 해답을 얻기 위해서는 그 分野의 專門知識이 요구되는 問題解決課題들이 많이 있다. 이러한 과제들은 잘 명시된 알고리즘에 의해서 그 解를 구할 수 없으며 대신 그 분야 專門家의 經驗的 知識과 판단력 (judgemental knowledge)에 의존하는 경우가 많다.

專門家시스템이나 知識基盤시스템과 같은 컴퓨터 프로그램은 人間專門家의 推論能力을 적용하기 위해서 개발되었다. 이러한 프로그램들이 기존의 프로그램들과 다른 큰 차이점은 問題領域에 관한 專門知識들을 굉장히 많이 이용하는데 있다. 이러한 지식들은 一般的인 推論方法에 의해 이용되고 숫자보다는 記號處理에 더 중점을 두며 또한 推論過程들은 자연스럽고 쉬운 方法으로 설명할 수 있는 기능이 있는 差異點이 있다. 專門家시스템은 지금까지 醫學診斷, 컴퓨터 構成, 部品故障診斷, 化學材料解析, 音聲處理 등에 성공적으로 적용되어 왔다.

지금까지의 일반적인 컴퓨터 問題解決方法은 정확한 수학적인 기반을 두고 있다. 즉, " yes " 또는 " no "를 결정할 수 있는 직접적인 意思決定能力에 국한되었다. 하지만 사람의 事故過程은 주로 부정확한 개념들(quantative formulation)을 다룬다. 예를 들면 "미끄러운 길에서 차를 빨리 모는 것은 매우 위험하다" "그 회로는 복잡하지만 신뢰도가 매우 높다"에서와 같이 밑줄친 단어들을 숫자로써 그 정도를 정확히 표시하기가 어려우며 또 그렇게 한다고 해도 부자연스러운 표현이 된다.

情報檢索시스템은 推論機能을 가지고 이와 같은 問題(부정확한, 어림치의 정보를 이용해서 추론)들을 다룰 수 있도록 구성되어야 하는 시스템의 한 예이다. 檢索시스템은 정확한 분석을 이용하기에 적합하지 못하다. 왜냐하면 각기 다른 사용자들이 각기 다른 시각을 가지고 시스템을 사용하기 때문이다. 따라서 이

시스템의 根本的인 特徵은 不正確性(imprecision)이다. 예를 들어 " very ", " similar ", " essentially ", " around "와 같은 단어들을 質疑語로 구성할 때 사용하는 것이 그 예이며 " recent ", " relevant ", " good ", " old "와 같은 단어들을 정확히 측정할 수 있는 質量的인 意味를 갖지 않고 있다. 또한 어떤 영역에서 중요한 용어가 여러개의 개념을 표현할 수 있다.

결론적으로 말해서 이러한 不正確性은 성격적으로 보아 統計的인 性質이 아니고 ambiguity의 문제이다. 이것은 정확한 質量的인 測定을 할 수 있는 方法의 不足으로 인한 것이다.

知識基盤시스템 接近方法은 개념들간의 포함관계 및 유사관계들을 명확히 명시하여 사용할 수 있으므로 文書檢索시스템에서 큰 효과를 얻을 수 있다. 이러한 接近方法은 사용자 질의어의 이해를 쉽게 할 수 있고 人間專門家와 비슷하게 적합한 문서들을 뽑아낼 수 있게 한다.

Ⅳ. 結　論

情報檢索分野에 人工知能의 技法들을 도입하여 많은 발전을 이루고 있고, 앞으로도 계속해서 개선되어 나가리라고 생각한다. 하지만 현재 상태에서 해결하지 못한 問題點들이 많으며 이러한 問題點들이 어떤 것인가를 명확히 파악하는 것도 큰 도움이 된다.

대부분의 情報檢索시스템이 궁극적으로는 人工知能의 큰 影響을 받게 될 것은 의심할 여지가 없다. 실제로 현재 개발되어 사용되는 DIALOG, ORBIT, MEDLARS, STAIRS와 같은 시스템들은 벌써 人工知能의 技法들을 많이 활용하고 있다. 이러한 시스템의 일부는 商業的으로 사용되고 있다. 하지만 이들은 一般性 (generality)과 融通性 (flexibility)이 부족하여 좁은 영역에서는 잘 동작되지만 넓은 영역에 대해 적용하려면 많은 問題點들이 노출된다.

결국 가장 큰 問題點은 실세계의 지식이 부족하다는 사실이다. 이러한 問題點들을 해결하는 것은 결국 현재 人工知能分野에서 해결하려는 問題點과 일치하게 된다.

따라서 情報檢索分野와 人工知能分野는 큰 연관관계를 맺으면서 발전되리라 생각한다.

〈參 考 文 獻〉

1. Michael Lebowitz, "Intelligent Information Systems," *Proceedings of Sixth ACM SIGIR Conference*, 1983, pp.5 ~ 30.
2. Gerald DeJong., "Artificial Intelligence Implications for Information Retrieval," *Proceedings of Sixth ACM SIGIR Conference*, 1983, pp.10 ~ 17.
3. Peretz Shoval., "Expert/Consultation System for a Retrieval Data-Base with Semantic Network of Concepts," *Proceedings of the ACM SIGIR Conference*, 1981, pp.145 ~ 150.
4. Madeleine Bates, Robert, and J.Bobrow, "Information Retrieval Using a Transportable Natural Language Interface," *Proceedings of the ACM SIGIR Conference*, 1983, pp.81 ~ 86.
5. Gian Piero ZARRI, "RESEDA, An Information Retrieval System Using Artificial Intelligence and Knowledge Representation Techniques," *Proceedings of ACM SIGIR Conference*, 1983, pp.189 ~ 195.
6. Roger H.Thompson and W.Bruce Croft, "An Expert System for Document Retrieval," *Expert Systems in Government Symposium*, 1985, pp.448 ~ 456.
7. Daniel A.Desalvo and Jay Liebowitz, "The Application of an Expert System for Information Retrieval At. the National Archieves," *Expert Systems in Government Symposium*, 1985, pp.464 ~ 472.
8. Robert R.Korflage, "Intelligent Information Retrieval : Issues in User Modelling," *Expert Systems in Government Symposium*, 1985, pp.474 ~ 482.
9. G.Biswas, V.Subramanian, and J.C.Beidek, "A Knowledge-Based System Approach to Document Retrieval," *Proceedings of the 2nd Conference on AI Applications*, December, 1985, pp.455 ~ 460.
10. S.G.Winett and E.A.Fox, "Using Information Retrieval Techniques in an

Expect System," *Proceedings of the 2nd Conference on AI Applications*, December, 1985, pp.230 ~ 235.

11. Gerard Salton, "A Blueprint for Automatic Indexing," *ACM SIGIR*, Vol. 16, No.2,(Fall 1981), pp.22~38.

12. Gerard Salton., "A Blueprint for Automatic Boolean Query Processing," *ACM SIGIR*, Fall, 1982, pp.6 ~ 24.

13. Gerard Salton., "Some Characteristics of Future Information Systems," *ACM SIGIR*, Fall, 1983, pp.28 ~ 39.

14. Linda C.Smith, "Artificial Intelligence in Information Retrieval Systems," *Information Processing & Management*, Vol.12, 1976, pp.189 ~ 222.

15. Gerard Salton and McGill, MJ., "Introduction to Modern Information Retrieval," McGraw-Hill, New York, 1983.

16. Gerard Salton and Edward A.Fox, Harry Wu. "Extended Boolean Information Retrieval," *Communications of the ACM*, 26, 1983, pp.1022 ~ 1036.

17. R.N.Oddy, "Information Retrieval through Man-Machine Dialogue," *Journal of Documentation*, Vol.33, No.1, March 1977, pp.1 ~ 14.

18. Tadeusz Radecki, "Similarity Measures for Boolean Search Request Formulations," *Journal of the American Society for Information Science*, January 1982, pp.8 ~ 17.

19. S.E.Robertson and K.Sparck Jones, "Relevance Weighting of Search Terms," *Journal of the American Society for Information Science*, May 1976, pp. 129 ~ 146.