情 報 檢 索

睦 延 均
〈KORSTIC 特許情報部長〉

—承 前—

1 情報檢索의 槪念

情報檢索(Information Retrieval=IR)은 情報管理分野에서 重要한 役割을 하고 있는 部門으로서 많은 說明과 技術이 必要하다.

本稿에서는 紙面關係로 核心的項目에 대해 要約記述키로 한다.

즉 情報檢索이라하면 情報의 發生源으로부터 利用者에 이르는 Communication System의 한 部分을 말하는 것으로서, 蒐集된 정보를 分析, 加工, 整理하여 어떤 形態로 蓄積(Storage)하여 두었다가 필요에 따라 검색(Retrieval)해 내게 되는 축적과 검색이라고 하는 두가지의 重要한 機能으로 되어 있다. 따라서 正確히 表現한다면 정보의 축적과 검색(Information Storage & Retrieval)이라 하겠다.

이와같이 정보의 검색을 위해서는 축적을 前提條件으로 하기 때문에 一般的으로 이를 줄여서 정보검색이라 하고 있는 데, 上記의 두가지 기능을 좀더 詳細히 설명하면 다음과 같다.

첫번째의 蓄積機能은 適時에 정보를 찾아내기 쉽도록하기 위해 수집되는 정보를 분석, 가공, 정리하여 蓄積處理를 하게 되는 데, 蓄積媒體로는 印刷物, 카드方式, Microfilm(MF)이나 Magnetic tape(MT)등에 收錄蓄積하는등 많은 方法이 있다. 여기서 MF나 MT와 같은 매체는 축적과 檢索過程에 撮影機, MF Reader, 電算機와 같은 機器가 필요하다. 그러나 蓄積狀態가 印刷媒體인 경우는 肉眼으로 直接볼 수 있는 長點을 지니고 있어서 가장 널리 活用되고 있다.

以上과 같은 축적처리를 하게되는 過程을 파일링(Filing) 또는 파일構成이라 한다. 이와 같은 蓄積處理工程을 볼 때 이를 구성하게되는 과정을 細分하여 보면 즉 수집, 정리, 파일구성의 과정을 거치게 된다.

다음으로 두번째의 검색기능은 축적파일의 情報中에서 要求에 따라 適合情報를 찾아내게되는 과정으로서 이를 이른바 探索(Searching)이라 하기도 한다. 탐색은 探索指令에 따라 축적된 파일에서 調査되어 回答情報를 얻게되는 것이다.

2 情報檢索의 區分

情報檢索은 蓄積情報의 內容, 檢索時點, 電算機에 의한 處理方法등 觀點에 따라 여러형태로 區分하게 된다.

1. 蓄積情報의 內容에 따른 區分

탐색코자하는 정보의 내용에 따라서 ① 데이터검색(Data Retrieval) 또는 事實檢索(Fact Retrieval) ② 文獻檢索(Document Retrieval)과 參照檢索(Reference Retrieval)등으로 구분하게 된다.

데이터검색이란 特定主題의 質問에 대해 필요한 數値情報, 즉 데이터 그 自體를 直接檢索하여내는 種類의 것으로, 경우에 따라서는 사실검색이라 하기도 한다.

그리고 문헌검색이란 利用者가 요구하는 原情報의 所在目錄 또는 抄錄과 같은 代用物을 검색하여 내게되는 것을 말한다. 이렇게 하여 關聯文獻의 所在가 하나 하나 羅列된 文獻目錄 또는 초록을 받게 되며, 要求者는 이들 목록이나 초록을 檢索한 후 필요한 것만 原文을 最終的으로 요구, 提供받게 된다. 여기서 단지 書誌(目錄)事項만을 檢索回答으로 받을 경우, 이를 참조검색

이라 말하기도 하는데 이를 엄격히 區別하여 使用하고 있지는 않다.

2. 探索時點을 내용으로한 區分

찾고자하는 요구정보를 탐색하는 시점을 기준하여 過去로 遡及하느냐 또는 새로히 入手되는 정보를 앞으로 追跡하여 탐색하느냐에 따라 ① 遡及探索(Rtrospective Search=RS) ② SDI (Selective bisseminatior of Information) 등의 두 형태로 구분한다.

상기의 遡及探索이란 探索時點을 基準으로 하여 過去分의 필요정보를 소급탐색하여 내게되는 것을 말하며,

SDI란 特定個人이나 團體가 필요로 하는 정보의 主題를 選定, 登錄하여 놓고 새로운 정보가 입수될 때마다 該當情報를 검색하여 定期的으로 제공하게 되는 것을 말하며, 이를 정보의 選擇提供, 最新情報의 現況追跡調査, 連續調査라는 등의 말로 사용되기도 한다.

3. 電算機의 處理方法에 의한 區分

電子計算機에 의한 정보검색에서는 정보의 처리방식 및 傳達方式에 따라 다음과 같이 구분한다.

處理方式 / 傳送方式	Batch 處理	Real—time處理
Off—line 方式	Off-line batch 處理	
On—line 方式	On-line batch 處理	On-line Real time 處理

몇해전만 하드라도 전산기에 의한 정보처리는 Off-line batch 처리였으나, 情報要求의 廣域化 및 即時化의 요구에 따라 現在는 On-line Real-time처리로 轉換되어 가고 있는 實情이다.

3 情報檢索結果의 評價

정보검색시스템에서 檢索結果를 評價하는 기준으로 다음과 같은 항목을 들 수 있다.

1. 檢索効率

정보검색시스템에 있어서 어떤目的의 정보를 검색할 경우 필요로하는 정보가 漏落됨이 없이 검색되고, 不要한 정보가 섞여있지 않게되면 理想的인 검색결과가 될 것이다. 그러나 이와같은 理想的인 검색결과를 얻는다는 것은 實際로 不可能하다. 그 理由로는 정보검색을 위한 수집된 정보의 主題分析, 索引파일作成, 蓄積 및 檢索過程에서 本意아닌 誤差(Error)가 發生하기 때문인 것이다.

이 오차를 統計學에서는 다음과 같이 설명하고 있다.

第1種의 誤差 : 假說이 옳바름에도 불구하고 이것을 棄却하는 잘못.

第2種의 誤差 : 가설이 옳바르지 못함에도 불구하고 이것을 採擇하는 잘못.

이 통계학의 오차를 정보검색에서 발생되는 오차에 應用하면

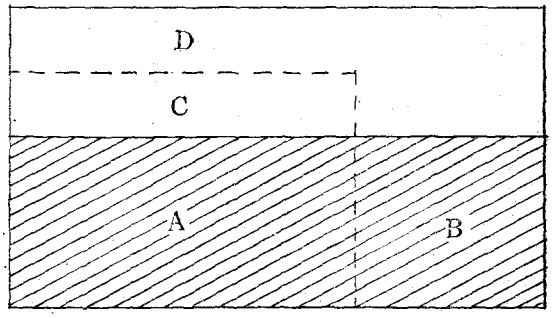
第1種의 誤差 : 어떤 引파일에 있는 適合情報를 一部만이 탐색되고 나머지를 탐색해내지 못할 경우

第2種의 誤差 : 탐색한 정보중에 不適合情報가 包含되어 있는 경우에 해당된다.

이와같이 정보의 탐색결과에 대한 効率을 測定하기 위한 尺度로서 精度(Precision) 또는 適合(Relevance)과 再現(Recall)이라고 한다.

정도 또는 적합이라고 하는 말은 탐색한 정보중에 利用者의 요구에 적합한 것이 얼마나 出力되었나를 나타내는 것이며, 재현이란 말은 어떤 색인파일에 있는 모든 關聯情報에서 탐색한 정보중 관련정보가 어느정도 抽出되었는가를 나타내는 것으로 이들을 百分率로 表示하여 精度率(Precision ratio) 또는 適合率(Relevance ratio)과 再現率(Recall ratio)이라 한다.

이를 그림을 통해 설명하면 다음과 같다.



어떤 색인파일에 축적되어 있는 정보의 總數를 直四角形으로 나타내고 質問에 대한 요구를 만족할만한 정보 즉 적합정보(Relevant information)가 一定量包含되어 있다고 假定하여 이를 편의상 斜線部分(A+B)으로 나타내기로 한다. 이와같은 검색과정에서 사선부분이 完全히 탐색되어진다면 이상적인 검색이 될 것이다. 그러나 實際的으로는 위에서 論한 第1種 및 第2種의 오차가 발생하여 결과적으로 點線으로 나타낸 C部分까지도 탐색되어 이 파일의 總體는 다음과 같이 될 것이다.

A : 탐색된 적합정보

A+B : 적합정보의 총수

B : 탐색되지 않은 적합정보(제 1종의 오차)

C : 탐색된 부적합정보(제 2종의 오차)

D : 탐색되지 않은 부적합정보

위에서 그림으로 볼때 精度(適合)率은 탐색된 全情報中의 적합정보의 百分率을 나타낸 것이고 재현율은 全適合情報中 探索된 적합정보의 百分率을 나타낸 것인데 이를 式으로 나타내면 다음과 같다.

$$\text{精度(適合)率} = \frac{\text{探索된 適合情報(A)}}{\text{探索된 全情報(A+C)}} \times 100$$

$$\text{再 現 率} = \frac{\text{探索된 適合情報(A)}}{\text{全 適合 情報(A+B)}} \times 100$$

또한 오차(Error)의 觀點으로 볼 때 전적합정보(A+B)중에서 탐색되지 않은 적합정보(B)의 百分率을 除外率(Omission ratio), 그리고 탐색된 전정보(A+C)에 대한 탐색된 부적합정보(C)의 百分率을 雜音率(Noise ratio)이라 한다.

2. 經濟性과 迅速性

어떠한 시스템을 設計할 경우일지라도 明確한 目標를 定하고 이들 목표를 될 수 있는대로 經濟的으로 迅速하게 達成토록 하여야 한다.

시스템의 經濟的評價에 있어서 먼저 정보의 價値, 機械化에 의한 有益性, 維持經費, 組織內에 있어서의 가치등을 고려함과 同時에 迅速性의 面에서도 滿足한 결과를 얻도록 하여야 할 것이다.

그러나 一般的으로 檢索效率, 經濟性, 迅速性은 兩立할 수 없는 경우가 大部分이다. 즉 효율좋은 검색결과를 迅速하게 얻기 위해서는 당연히 高價의 費用이 요구되기 때문이다.

그러므로 각 시스템에 따라 最適基準을 정하고 運營하여 나아가야 할 것이다.

4 情報檢索方式

情報檢索方式에는 매뉴얼시스템(Manual System)과 기계화시스템으로 구분하게 된다.

1. 매뉴얼시스템

매뉴얼시스템이란 自動化된 機械裝置에 의하지 않고 카드와 단순한 手動式器具만을 사용하여 수동으로 정보를 축적하고 검색하는 방법을 말하는데 다음과 같은 방식들이 있다.

(1) 單純카드시스템(Plane Card System)

(2) Uniterm Card System

(3) Pee-a-boo Card System

(4) 手動式펀치카드시스템(Hanb Sorteb Punched Card System)

2. 機械化시스템

정보검색의 효율화를 위하여 各種機器가 開發되어 이용되고 있다. 즉 手作業을 代行하는 단순한 機器로부터 人間頭腦의 役割을 代身하는 電子計算機에 이르기까지 많은 종류의 情報檢索機器들이 활용되고 있는데, 이들에 대한 效率的인 활용을 위해서는 企業의 실정과 利用目的에 따라 알맞는 시스템을 導入하여야 할 것이다.

기계화시스템에는 다음과 같은 방식들이 있다.

(1) 多欄 Carb Selector

(2) 마이크로필름을 이용한 검색시스템

① Aperture Card System

② Image Control System

③ MIRA Code System

④ Filmorex System

⑤ 기타 마이크로필름검색시스템

(3) Vibeo file System

(4) 전자계산기를 이용한 검색시스템

〈完〉